Research Article

Statistics in Medicine

Received 13 September 2011, Accepted 22 February 2012 Published online 22 June 2012 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5372

Adaptive extensions of a two-stage group sequential procedure for testing primary and secondary endpoints (I): unknown correlation between the endpoints

Ajit C. Tamhane,[a*†] Yi Wu[b] and Cyrus R. Mehta[c]

In a previous paper we studied a two-stage group sequential procedure (GSP) for testing primary and secondary endpoints where the primary endpoint serves as a gatekeeper for the secondary endpoint. We assumed a simple setup of a bivariate normal distribution for the two endpoints with the correlation coefficient ρ between them being either an unknown nuisance parameter or a known constant. Under the former assumption, we used the least favorable value of $\rho = 1$ to compute the critical boundaries of a conservative GSP. Under the latter assumption, we computed the critical boundaries of an exact GSP. However, neither assumption is very practical. The $\rho = 1$ assumption is too conservative resulting in loss of power, whereas the known ρ assumption is never true in practice. In this part I of a two-part paper on adaptive extensions of this two-stage procedure (part II deals with sample size re-estimation), we propose an intermediate approach that uses the sample correlation coefficient r from the first-stage data to adaptively adjust the secondary boundary after accounting for the sampling error in r via an upper confidence limit on ρ by using a method due to Berger and Boos. We show via simulation that this approach achieves 5–11% absolute secondary power gain for $\rho \leqslant 0.5$. The preferred boundary combination in terms of high primary as well as secondary power is that of O'Brien and Fleming for the primary and of Pocock for the secondary. The proposed approach using this boundary combination achieves 72–84% relative secondary power gain (with respect to the exact GSP that assumes known ρ). We give a clinical trial example to illustrate the proposed procedure. Copyright © 2012 John Wiley & Sons, Ltd.

Keywords: adaptive designs; familywise error rate; gatekeeping procedures; multiple comparisons; O'Brien–Fleming boundary; Pocock boundary

1. Introduction

Tamhane *et al.* [1] studied a two-stage group sequential procedure (GSP) for the problem of testing primary and secondary endpoints with the former acting as a gatekeeper for the latter. Hung *et al.* [2] were the first to study this problem. They compared three different strategies via simulation in terms of the type I error rate control for different values of ρ. They showed that the strategy that tests the secondary null hypothesis at level α upon rejecting the primary null hypothesis does not control the error rate, but the strategy that tests the secondary null hypothesis at level $\alpha/2$ is conservative. Finally, the strategy that tests both the primary and the secondary null hypotheses by using the same α-level critical boundary controls the error rate more accurately. Glimm *et al.* [3] independently obtained the same results as in [1]; they also considered other variations of the problem such as two coprimary endpoints instead of one primary and one secondary.

Group sequential procedures have been around for more than 40 years beginning with the works by Armitage ([4, 5]). Pocock [6] proposed constant critical boundaries for GSPs, whereas O'Brien and Fleming [7] proposed decreasing critical boundaries. Lan and DeMets [8] proposed the use of error rate

[a] *Department of IEMS, Northwestern University, Evanston, IL 60208, U.S.A.*
[b] *Department of Statistics, Northwestern University, Evanston, IL 60208, U.S.A.*
[c] *Cytel Corporation and Harvard School of Public Health, Cambridge, MA 02139, U.S.A.*
**Correspondence to: Ajit C. Tamhane, Department of IEMS, Northwestern University, Evanston, IL 60208, U.S.A.*
[†] *E-mail: atamhane@northwestern.edu*

spending functions, which can be chosen to design the critical boundaries to suit the needs of the trial. Jennison and Turnbull's book [9] is a comprehensive reference on the topic.

Tamhane *et al.* [1] assumed that the correlation coefficient ρ between the primary and secondary endpoints is either an unknown nuisance parameter or a known constant. Under the former assumption, they showed that $\rho = 1$ is the least favorable value of ρ and used it to compute the critical boundaries of the GSP. They also computed the critical boundaries of the exact GSP for selected known values of ρ. However, neither assumption is very practical. The $\rho = 1$ assumption is too conservative resulting in loss of power, whereas the known ρ assumption is never true in practice. In this paper, we show how to use the sample correlation coefficient r from the first-stage data in place of ρ to adaptively adjust the secondary boundary. The sampling error in r is taken into account via an upper confidence limit on ρ following an approach due to Berger and Boos [10]. We show that this method achieves substantial power gains: 5–11% absolute power gain and 72–84% relative power gain (with respect to the exact GSP that assumes known ρ) compared with the conservative method that assumes $\rho = 1$.

We organize the paper as follows. Section 2 defines the notation and gives a brief review of the Tamhane *et al.* [1] procedure. Section 3 presents the confidence limit method to deal with unknown correlation. Section 4 gives the details of the calculation of the optimum critical boundary for the secondary endpoint using the confidence limit method. Section 5 discusses a simulation study to study the robustness of the proposed procedure in terms of control of the familywise error rate (FWER) when the variances of the primary and secondary endpoints are estimated from the data instead of being known as assumed for convenience in the paper. Section 6 gives the results of the secondary power comparisons of the proposed method with the conservative method and the exact method. In Section 7, we give the modifications needed to extend the methodology developed in this paper for a single sample case to the two-sample case (matched pairs or independent samples) necessary to deal with two parallel arm trials. Section 8 gives an illustrative clinical trial example. Section 9 discusses the problem involved in extending the methodology to binary data and gives some concluding remarks.

2. Notation and background

Assume a two-stage GSP with sample sizes n_1 and n_2. For the sake of simplicity, we will consider the single sample case. Let (X_{ij}, Y_{ij}) be independent and identically distributed bivariate normal observations on the primary and secondary endpoints for the jth patient in the ith stage $(i = 1, 2, j = 1, \ldots, n_i)$, where $X_{ij} \sim N(\mu_1, \sigma_1^2)$, $Y_{ij} \sim N(\mu_2, \sigma_2^2)$, and $\mathrm{corr}(X_{ij}, Y_{ij}) = \rho \geqslant 0$. Let $\delta_1 = \mu_1/\sigma_1$ and $\delta_2 = \mu_2/\sigma_2$. All parameters are unknown except σ_1 and σ_2, which are assumed to be known for convenience. The hypotheses to be tested are $H_1 : \delta_1 = 0$ and $H_2 : \delta_2 = 0$ against upper one-sided alternatives, subject to the gatekeeping restriction that H_2 can be tested iff H_1 is rejected; otherwise H_2 is accepted without testing it.

The first-stage test statistics are defined as

$$X_1 = \frac{\overline{X}_1}{\sigma_1/\sqrt{n_1}} \text{ and } Y_1 = \frac{\overline{Y}_1}{\sigma_2/\sqrt{n_1}}, \tag{1}$$

where $\overline{X}_1 = \sum_{j=1}^{n_1} X_{1j}/n_1$ and $\overline{Y}_1 = \sum_{j=1}^{n_1} Y_{1j}/n_1$ are the first-stage sample means. Similarly, the second-stage test statistics are defined as

$$X_2 = \frac{\overline{X}_2}{\sigma_1/\sqrt{n_1+n_2}}, Y_2 = \frac{\overline{Y}_2}{\sigma_2/\sqrt{n_1+n_2}}, \tag{2}$$

where $\overline{X}_2$ and $\overline{Y}_2$ are the overall sample means.

The joint distribution of (X_1, Y_1, X_2, Y_2) is four-variate normal with the following means:

$$E(X_1) = \delta_1\sqrt{n_1}, E(Y_1) = \delta_2\sqrt{n_1}, E(X_2) = \delta_1\sqrt{n_1+n_2}, E(Y_2) = \delta_2\sqrt{n_1+n_2}. \tag{3}$$

In the following, we denote $\Delta_1 = \delta_1\sqrt{n_1}$ and $\Delta_2 = \delta_2\sqrt{n_1}$. Let $f = n_1/(n_1 + n_2)$ be the information fraction [5] at the interim look . The correlation structure of (X_1, Y_1, X_2, Y_2) is given by

$$\begin{aligned} \mathrm{corr}(X_1, Y_1) &= \mathrm{corr}(X_2, Y_2) = \rho \\ \mathrm{corr}(X_1, X_2) &= \mathrm{corr}(Y_1, Y_2) = \tau \\ \mathrm{corr}(X_1, Y_2) &= \mathrm{corr}(X_2, Y_1) = \rho\tau, \end{aligned} \tag{4}$$

where $\tau = \sqrt{f}$. Denote the primary critical boundary for (X_1, X_2) by (c_1, c_2) and the secondary critical boundary for (Y_1, Y_2) by (d_1, d_2). The GSP, denoted by $\mathcal{P}$, operates as follows.

Stage 1. Take n_1 observations, $(X_{1j}, Y_{1j}), j = 1, \ldots, n_1$, 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, stop sampling.

Stage 2. Take n_2 observations, $(X_{2j}, Y_{2j}), j = 1, \ldots, n_2$, 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.

The critical boundaries (c_1, c_2) and (d_1, d_2) of $\mathcal{P}$ must be determined to satisfy the following FWER control requirement:

$$\text{FWER} = P\{\text{Reject at least one true} \, H_i \, (i = 1, 2)\} \leqslant \alpha \tag{5}$$

for a specified α when either H_1 or H_2 is true.

Tamhane *et al.* [1] showed that FWER is controlled at level α under H_1 if (c_1, c_2) is an α-level boundary, that is,

$$P_{H_1}(X_1 > c_1) + P_{H_1}(X_1 \leqslant c_1, X_2 > c_2) = \alpha. \tag{6}$$

For example, we can use the O'Brien–Fleming (OF) [7] boundary, which uses $c_1 = c\sqrt{2}, c_2 = c$, or the Pocock (PO) [6] boundary, which uses $c_1 = c_2 = c$, where $c > 0$ is determined in each case to satisfy (6).

Under H_2, FWER is a function of Δ_1 and ρ (denoted by FWER(Δ_1, ρ)). To control FWER for given (c_1, c_2), we need to determine (d_1, d_2) so that $\max_{\Delta_1,\rho} \text{FWER}(\Delta_1, \rho) \leqslant \alpha$. It was shown in [1] using numerical methods that $\max_{\Delta_1} \text{FWER}(\Delta_1, \rho)$ is an increasing function of ρ. Furthermore, it was shown analytically that the overall maximum of FWER(Δ_1, ρ) occurs when $\rho = 1$ and $\Delta_1 = \Delta_1^*$ where Δ_1^* depends on (c_1, c_2) and (d_1, d_2). In particular, if both (c_1, c_2) and (d_1, d_2) are α-level boundaries as defined in (6) and $c_1 \geqslant d_1$ and $c_2 \leqslant d_2$ (e.g., if (c_1, c_2) is the OF boundary and (d_1, d_2) is the PO boundary in which case $c_1 > d_1$ and $c_2 < d_2$ or if $(c_1, c_2) = (d_1, d_2)$) then $\max_{\Delta_1,\rho} \text{FWER}(\Delta_1, \rho) = \text{FWER}(\Delta_1^*, 1) = \alpha$ where $\Delta_1^* = c_1 - d_1$. On the other hand, if $c_1 < d_1$ 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 $\max_{\Delta_1,\rho} \text{FWER}(\Delta_1, \rho) = \text{FWER}(\Delta_1^*, 1) < \alpha$, where $\Delta_1^* = \tau(c_2 - d_2)$; therefore, the α-level secondary boundary (d_1, d_2) can be changed to an α'-level secondary boundary (d_1', d_2'), where $\alpha' > \alpha$ to make $\max_{\Delta_1,\rho} \text{FWER}(\Delta_1, \rho) = \alpha$.

Tamhane *et al.* [1] compared the secondary powers (i.e., the powers to reject false H_2) of different combinations of primary and secondary boundaries, for example, OF1–OF2, OF1–PO2, PO1–OF2, and PO1–PO2 under various alternatives. They found that the OF1–PO2 combination is the most powerful among all four combinations except when Δ_1 and Δ_2 are both small. Note also that the primary power (i.e., the power to reject false H_1) for the OF1 boundary is uniformly (for all Δ_1) higher than that for the PO1 boundary.

3. Confidence limit method

As can be seen from the previous text, the FWER control requirement (5) can be satisfied by choosing both (c_1, c_2) and (d_1, d_2) to be α-level boundaries, that is, they satisfy (6). However, the α-level (d_1, d_2) boundary can be overly conservative because it assumes the least favorable value $\rho = 1$. Because the true ρ is unknown, an important practical problem is how to choose a less conservative (d_1, d_2) boundary by using the sample correlation coefficient r from the first-stage data. If (d_1, d_2) are determined simply by substituting r for the unknown ρ, then FWER $> \alpha$ if $r < \rho$ (because the (d_1, d_2) boundary is underestimated) and FWER $< \alpha$ if $r > \rho$ (because the (d_1, d_2) boundary is overestimated). Thus, even though the average FWER may be close to the nominal α, it can exceed α in a significant proportion of cases. Figure 1 shows the plot of simulated proportion of times the FWER exceeds the nominal $\alpha = 0.05$ as a function of Δ_1 when the sample estimate r is used in place of ρ; here, the true ρ is set equal to 0.5 and $n_1 = 20$. We see that for $\Delta_1 \approx 1.5$, the FWER exceeds $\alpha = 0.05$ in about 50% of the cases. Therefore, we should not simply substitute r for the true ρ. To deal with this nuisance parameter problem, we follow the approach of Berger and Boos [10].

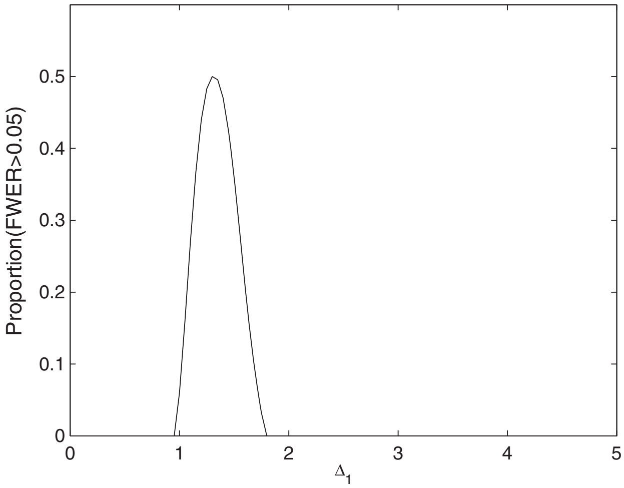

Figure 1. Proportion of simulation runs of $\mathcal{P}$ in which FWER > 0.05 if r is used as the true ρ when $\rho = 0.5$ and $n_1 = 20$.

Let ρ^* be a $100(1-\varepsilon)\%$ upper confidence limit on ρ, that is, $P(\rho \leqslant \rho^*) = 1-\varepsilon$. To calculate ρ^*, we used Fisher's arctan hyperbolic transformation:

$$\frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) \sim_{\text{approx.}} N\left(\frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right), \frac{1}{n_1-3}\right), \tag{7}$$

which leads to the following approximate $100(1-\varepsilon)\%$ upper confidence limit:

$$\rho^* = \frac{e^{2u}-1}{e^{2u}+1}, \text{ where } u = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) + \frac{z_\varepsilon}{\sqrt{n_1-3}}, \tag{8}$$

where z_ε is the $100(1-\varepsilon)$ percentile of the standard normal distribution. Using the property that $\max_{\Delta_1}$ FWER(Δ_1, ρ) is an increasing function of ρ, we can derive an upper bound on this maximum for any unknown ρ as follows: let $\Delta_1^*(\rho)$ be the value of Δ_1 that maximizes FWER(Δ_1, ρ) for fixed ρ. Then,

$$\begin{aligned}\max_{\Delta_1} \text{FWER}(\Delta_1,\rho) &\leqslant \max_{\{\rho\leqslant\rho^*\}} \text{FWER}(\Delta_1^*(\rho),\rho)\times P(\rho\leqslant\rho^*) + \max_{\{\rho>\rho^*\}} \text{FWER}(\Delta_1^*(\rho),\rho)\times P(\rho>\rho^*)\\ &= \text{FWER}(\Delta_1^*(\rho^*),\rho^*)\times(1-\varepsilon) + \text{FWER}(\Delta_1^*(1),1)\times\varepsilon\\ &= \alpha'(1-\varepsilon)+\alpha''\varepsilon,\end{aligned} \tag{9}$$

where $\alpha' = \text{FWER}(\Delta_1^*(\rho^*),\rho^*) < \alpha'' = \text{FWER}(\Delta_1^*(1),1)$ are functions of (d_1, d_2).

We want to determine the sharpest possible (d_1, d_2) (so as to maximize the secondary power) subject to (9) $\leqslant \alpha$. This problem can be solved numerically on a computer as follows. Suppose the boundary (d_1, d_2) is parameterized through some common d, for example, $d_1 = d\sqrt{2}$, $d_2 = d$ for the OF2 boundary and $d_1 = d_2 = d$ for the PO2 boundary. Then, we choose the optimum confidence level $1-\varepsilon$ so as to minimize d and thus maximize the secondary power. First, note that if $1-\varepsilon$ is increased, then ρ^* increases causing α' to increase while α'' is fixed. Also, the weight $1-\varepsilon$ on α' increases while the weight ε on α'' decreases. The net result is that as $1-\varepsilon$ increases, the upper bound in (9) first decreases (because the weight ε on $\alpha'' > \alpha'$ decreases) and then increases. We should then choose, for each given n_1 and r, the value of $1-\varepsilon$ that minimizes the overall max FWER and choose d to make this minimax FWER equal to α. This is in essence the confidence limit method.

Remark 1

It should be noted that because ρ^* is a random variable, the FWER control guarantee derived in (9) is conditional on observed r. However, this guarantee holds for every observed r and so it holds unconditionally.

4. Calculation of the optimum (d_1, d_2) boundary

To calculate the optimum (d_1, d_2) boundary for given sample correlation coefficient r, the primary boundary (c_1, c_2), and the sample size $n_1 = n_2 = n$, we considered four cases: the (c_1, c_2) boundary is either OF (in which case $c_1 = c_2\sqrt{2}$) or PO (in which case $c_1 = c_2$). For each choice of the primary boundary, we considered the same two choices for the secondary boundary: OF (in which case $d_1 = \sqrt{2}d, d_2 = d$) or PO (in which case $d_1 = d_2 = d$). Table I gives the optimum values of d for observed $r = 0.1(0.1)1.0$, $n_1 = n_2 = n = 20, 50, 100$ and $\alpha = 0.05$. For comparison purposes, we have also included the corresponding values of d for known ρ (i.e., $n_1 = \infty$), which are, of course, smaller. Note that Table I also lists the associated confidence coefficient $1 - \varepsilon$ (from which the upper confidence limit ρ^* can be computed using (8)). However, to implement the procedure, the intermediate quantities ρ^* and associated $1 - \varepsilon$ are not needed; only the secondary boundary (d_1, d_2).

From Table I we see that, as expected, for any given combination of OF and PO boundaries and for any given r, as n increases d decreases, approaching the limit for known ρ as $n \to \infty$. For given n, as r increases, the confidence coefficient $1 - \varepsilon$ decreases and for given r, as n increases, the confidence coefficient $1 - \varepsilon$ increases. The explanation for this behavior is as follows. As r increases, the upper confidence limit ρ^* becomes close to 1 if $1 - \varepsilon$ becomes large, which makes the first term in (9) large, whereas the second term only decreases slightly because of the decrease in ε because $\mathrm{FWER}(\Delta_1^*(1), 1)$ is fixed. Hence, to compensate for the increase in ρ^* as a result of the increase in r and the consequent increase in $\mathrm{FWER}(\Delta_1^*(\rho^*), \rho^*)$, the confidence coefficient $1 - \varepsilon$ must decrease.

Table I. The optimum d-values for the secondary boundary ($d_1 = \sqrt{2}d, d_2 = d$ for OF2, $d_1 = d_2 = d$ for PO2) and the associated confidence level $1 - \varepsilon(n_1 = n_2 = n, \alpha = 0.05)$.

		Observed r									
Procedure	n	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
OF1–OF2	20	1.489	1.503	1.517	1.532	1.547	1.565	1.586	1.611	1.635	1.678
		(0.94)	(0.93)	(0.92)	(0.91)	(0.89)	(0.86)	(0.79)	(0.69)	(0.64)	
	50	1.463	1.477	1.492	1.507	1.524	1.543	1.564	1.590	1.622	1.678
		(0.97)	(0.97)	(0.96)	(0.96)	(0.95)	(0.94)	(0.92)	(0.87)	(0.81)	
	100	1.450	1.464	1.478	1.494	1.512	1.531	1.553	1.579	1.612	1.678
		(0.98)	(0.98)	(0.98)	(0.97)	(0.97)	(0.97)	(0.96)	(0.96)	(0.96)	
	∞	1.416	1.428	1.440	1.455	1.473	1.493	1.519	1.551	1.591	1.678
OF1–PO2	20	1.713	1.724	1.735	1.746	1.758	1.771	1.786	1.805	1.832	1.876
		(0.95)	(0.94)	(0.93)	(0.93)	(0.92)	(0.91)	(0.89)	(0.81)	(0.65)	
	50	1.692	1.703	1.715	1.727	1.740	1.755	1.771	1.791	1.822	1.876
		(0.97)	(0.97)	(0.96)	(0.96)	(0.96)	(0.95)	(0.94)	(0.91)	(0.83)	
	100	1.681	1.692	1.704	1.716	1.730	1.745	1.762	1.783	1.811	1.876
		(0.98)	(0.98)	(0.98)	(0.98)	(0.97)	(0.97)	(0.97)	(0.96)	(0.91)	
	∞	1.652	1.663	1.673	1.686	1.699	1.717	1.735	1.760	1.791	1.876
PO1–OF2	20	1.368	1.383	1.397	1.412	1.429	1.447	1.465	1.490	1.516	1.570
		(0.94)	(0.92)	(0.91)	(0.90)	(0.87)	(0.84)	(0.83)	(0.76)	(0.71)	
	50	1.341	1.355	1.370	1.387	1.407	1.426	1.447	1.471	1.505	1.570
		(0.97)	(0.97)	(0.96)	(0.94)	(0.91)	(0.89)	(0.88)	(0.87)	(0.79)	
	100	1.327	1.341	1.356	1.372	1.391	1.410	1.434	1.459	1.494	1.570
		(0.98)	(0.98)	(0.97)	(0.97)	(0.96)	(0.96)	(0.94)	(0.93)	(0.91)	
	∞	1.290	1.304	1.317	1.333	1.350	1.372	1.396	1.429	1.470	1.570
PO1–PO2	20	1.697	1.707	1.717	1.729	1.741	1.756	1.774	1.793	1.818	1.876
		(0.96)	(0.95)	(0.95)	(0.93)	(0.91)	(0.89)	(0.85)	(0.84)	(0.80)	
	50	1.678	1.687	1.697	1.709	1.722	1.739	1.759	1.781	1.809	1.876
		(0.98)	(0.98)	(0.97)	(0.96)	(0.96)	(0.92)	(0.88)	(0.85)	(0.82)	
	100	1.669	1.677	1.687	1.699	1.712	1.727	1.745	1.768	1.802	1.876
		(0.99)	(0.99)	(0.98)	(0.98)	(0.97)	(0.96)	(0.95)	(0.94)	(0.89)	
	∞	1.648	1.655	1.661	1.672	1.683	1.698	1.716	1.742	1.777	1.876

The parenthetical entry below each optimum d is the corresponding confidence coefficient $1 - \varepsilon$.

5. Unknown variances

Thus far, we assumed that the variances, σ_1^2 and σ_2^2, of the primary and secondary endpoints are known and the critical boundaries (c_1, c_2) and (d_1, d_2) were computed under this assumption. In practice, one needs to use the sample estimates of σ_1^2 and σ_2^2 to compute the first-stage and second-stage test statistics. To check whether the FWER is controlled under this setting, we conducted a small simulation study. We focused on FWER control under the secondary null hypothesis, $H_2 : \mu_2 = 0$, because the error rate control under the primary null hypothesis when using an α-level primary boundary (c_1, c_2) is guaranteed and does not involve the confidence limit method. Specifically, we simulated $X_{ij} \sim N(\delta_1, 1)$, $Y_{ij} \sim N(0, 1)$ and $\mathrm{corr}(X_{ij}, Y_{ij}) = \rho$ for $i = 1, 2$, $j = 1, \ldots, n_i$ with δ_1 equal to $\Delta_1^*(\rho)/\sqrt{n_1}$, where $\Delta_1^*(\rho)$ is the value of Δ_1 that maximizes the FWER for given ρ. We chose $\rho = 0.5$ and $n_1 = n_2 = n = 20, 25, 50,$ and 100.

For each value of n, we simulated the proposed procedure 10,000 times using the OF1–PO2 boundaries. For each simulation, we generated the first-stage sample from which we calculated the estimates of (σ_1^2, σ_2^2) and then (X_1, Y_1) and r. If the procedure did not stop at the first stage (i.e., if $X_1 \leqslant c_1$), then we computed the boundary (d_1, d_2) with $d_1 = d_2 = d$ for the observed r using the confidence limit method (which sets $\max_{\Delta_1} \mathrm{FWER}(\Delta_1, \rho)$ as close to α as possible). Next, we generated the second-stage sample, calculated the pooled (from both stages) estimates of (σ_1^2, σ_2^2), and then (X_2, Y_2). Finally, we tested X_2 against c_2, and if H_1 is rejected, then tested Y_2 against d_2 to see if H_2 can be rejected. Table II shows the estimates of FWER obtained.

From the table, we see that the achieved FWER is very close to the nominal $\alpha = 0.05$ when the variances are estimated even for n as small as 20. In practice, typically $n \geqslant 50$ is used. Also note that these simulations are made under the least favorable value of Δ_1, and so in other cases, the achieved FWER will be even less. Therefore, using the sample variances in place of the unknown variances is not a problem.

6. Power comparisons

We carried out two separate secondary power comparisons, both focused on assessing the advantage of the confidence limit method over the conservative method. (Note that the primary power to reject H_1 depends on the primary boundary (c_1, c_2) and Δ_1 and is not influenced by the secondary endpoint.) The exact method was also included in the comparisons as the gold standard. Power comparison I compared the three methods using four different combinations of OF and PO primary and secondary boundaries for different values of ρ for fixed Δ_1 and Δ_2. Power comparison II compared the three methods for $\rho = 0.3, 0.5, 0.7$ and fixed Δ_1 for different values of Δ_2, where Δ_1 was chosen to guarantee 80% primary power. In this case, only the OF1–PO2 boundary combination was studied because power comparison I showed that it dominates the other three combinations in terms of the secondary power for all three methods, in particular for the confidence limit method.

6.1. Power comparison I

We carried out power computations for $\rho = 0.1(0.1)1.0$, $\Delta_1 = 3$, $\Delta_2 = 2$, $n_1 = n_2 = n = 20, 50, 100$, and for four different combinations of OF and PO primary and secondary boundaries. We computed the secondary powers for the conservative and the exact methods by using the integral expression (8) in [1]. In the case of the confidence limit method, we used simulations because ρ^* is a random variable. For given ρ and n_1, we generated 10,000 values of the sample correlation coefficient r from its approximate distribution given by (7). For each realization of r, we calculated the optimum value of d by interpolating in Table I; we obtained the optimum (d_1, d_2) boundary from this d as explained before. We then calculated the secondary power for this optimum (d_1, d_2) using the aforementioned integral expression. Finally, we used the average of the secondary powers thus calculated for the 10,000 simulated values of r as an estimate of the true secondary power of this method. In Table III, we report the results for $n = 50$.

Table II. Estimated FWER when variances are estimated for selected $n_1 = n_2 = n$ values ($\alpha = 0.05$).

n	20	25	50	100
FWER	0.0491	0.0496	0.0498	0.0491

Table III. Secondary power comparison between the confidence limit method, the conservative method using $\rho = 1$ and the known true ρ method ($\Delta_1 = 3$, $\Delta_2 = 2$, $n_1 = n_2 = 50, \alpha = 0.05$).

		True ρ									
Procedure	Method (ρ)	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1.0
OF1–OF2	$\rho =$ True ρ	0.6162	0.6204	0.6244	0.6269	0.6275	0.6266	0.6212	0.6099	0.5877	0.5254
	$\rho = \rho^*$	0.5946	0.5975	0.5996	0.6013	0.6015	0.6000	0.5960	0.5866	0.5679	0.5254
	$\rho = 1$	0.4958	0.5029	0.5097	0.5161	0.5220	0.5273	0.5315	0.5338	0.5322	0.5254
OF1–PO2	$\rho =$ True ρ	0.7045	0.7091	0.7142	0.7186	0.7234	0.7268	0.7308	0.7327	0.7324	0.6945
	$\rho = \rho^*$	0.6912	0.6956	0.6999	0.7043	0.7087	0.7128	0.7169	0.7200	0.7183	0.6945
	$\rho = 1$	0.6270	0.6344	0.6420	0.6497	0.6576	0.6659	0.6746	0.6837	0.6931	0.6945
PO1–OF2	$\rho =$ True ρ	0.6173	0.6171	0.6172	0.6156	0.6129	0.6067	0.5977	0.5807	0.5540	0.4861
	$\rho = \rho^*$	0.5917	0.5912	0.5899	0.5872	0.5822	0.5767	0.5684	0.5556	0.5324	0.4861
	$\rho = 1$	0.4747	0.4801	0.4850	0.4894	0.4931	0.4958	0.4971	0.4963	0.4924	0.4861
PO1–PO2	$\rho =$ True ρ	0.6680	0.6712	0.6748	0.6766	0.6783	0.6782	0.6762	0.6691	0.6538	0.6020
	$\rho = \rho^*$	0.6574	0.6598	0.6618	0.6630	0.6637	0.6623	0.6589	0.6526	0.6395	0.6020
	$\rho = 1$	0.5846	0.5896	0.5945	0.5992	0.6036	0.6075	0.6104	0.6115	0.6091	0.6020

For each boundary combination, the table entries are the secondary powers corresponding to the (d_1, d_2) boundary computed by three methods. Top row: known ρ method; middle row: upper confidence limit method; and bottom row: conservative method.

By examining this table, we can see that for each one of the three methods and for every value of the true ρ, the OF1–PO2 boundary combination is more powerful than the other three boundary combinations. Recall that the OF1–PO2 combination was also found to be generally more powerful than other boundary combinations for the conservative method in [1].

To get a clearer picture of the differences between the three methods, we have made the plots of the secondary powers versus the true ρ in Figures 2 and 3. We see that when the true ρ is small, the absolute power gains of the proposed method with respect to the conservative method are high ranging from 9% to 11% for OF1–OF2 and PO1–OF2 boundary combinations and 5% to 7% for OF1–PO2 and PO1–PO2 boundary combinations for $\rho \leqslant 0.5$. Although it is difficult to quantify precisely what an 11% gain in secondary power implies for sample size (which is normally determined on the basis of the primary power considerations), it is instructive to note that the sample size needed to boost primary power by 11% can be substantial. For example, if we were designing a two-sample t-test for 80% primary power with a two-sided 0.05-level test and 50 subjects per arm, the same test would require 70 subjects per arm for 91% power, a 40% increase in sample size.

As a further graphical aid for comparing the secondary powers of the three methods, we have plotted

$$\text{Power Gain}(\%) = \frac{\text{Power } (\rho = \rho^*) - \text{Power } (\rho = 1)}{\text{Power (Known } \rho) - \text{Power } (\rho = 1)} \times 100 \tag{10}$$

as a function of ρ in Figure 4 for the OF1–PO2 boundary combination. Note that the percentage power gain decreases as ρ increases. This is explained readily because as ρ increases to 1, all three methods converge.

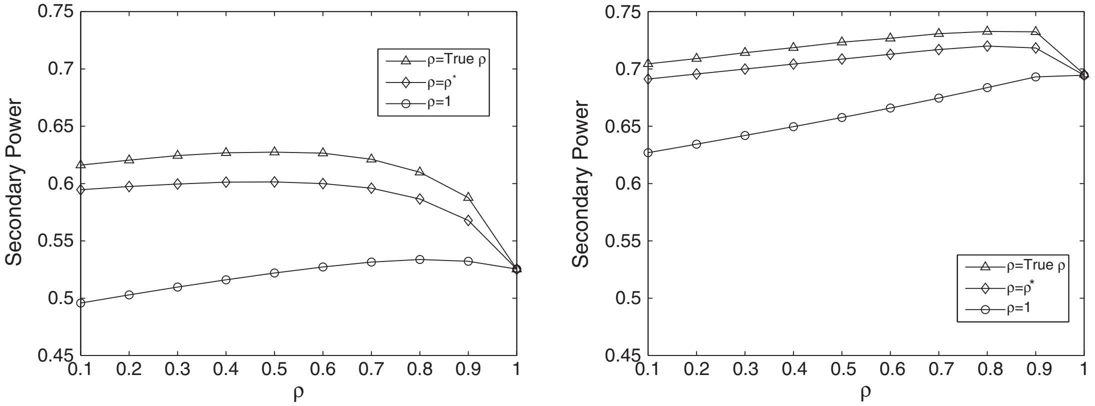

Figure 2. Plots of secondary powers (left panel: OF1–OF2 boundary combination; right panel: OF1–PO2 boundary combination) as functions of true ρ for the exact method, confidence limit method, and conservative method.

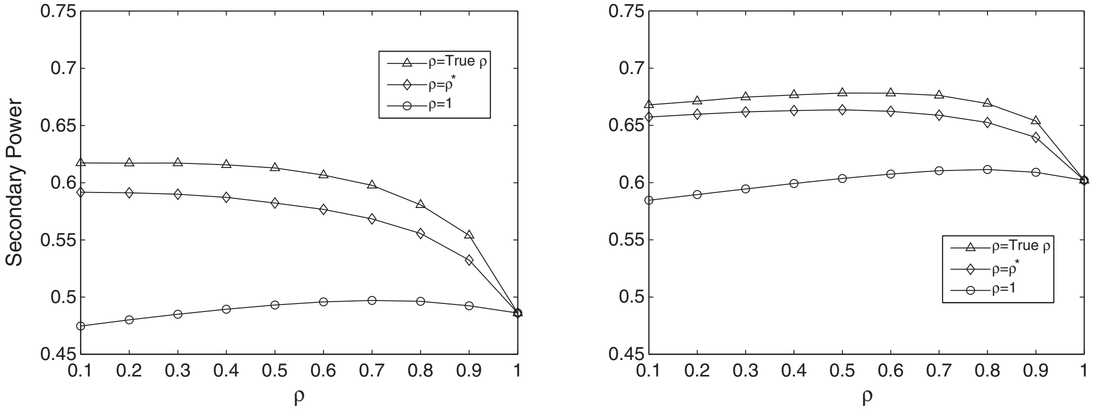

Figure 3. Plots of secondary powers (left panel: PO1–OF2 boundary combination; right panel: PO1–PO2 boundary combination) as functions of true ρ for the exact method, confidence limit method, and conservative method.

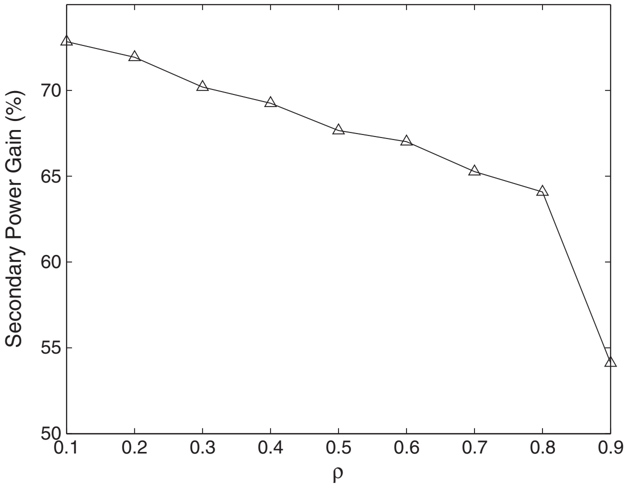

Figure 4. Power gain (%) of the confidence limit method as a function of ρ for OF1–PO2 boundary combination.

6.2. *Power comparison II*

We made the aforementioned secondary power comparisons for fixed $\Delta_1 = 3$ and $\Delta_2 = 2$. To assess the power gains as a function of Δ_2, we considered the following setup. Fix $\alpha = 0.05$ and the primary power $1-\beta = 0.80$. Then, for the OF primary boundary with two stages, the common sample size n per stage is given by

$$n = R\left(\frac{z_\alpha + z_\beta}{\delta_1}\right)^2 = (1.016)\left(\frac{1.645 + 0.840}{\delta_1}\right)^2,$$

where $R = 1.016$ is taken from [9, Table 2.4]. (Because we are considering a one-sided test with $\alpha = 0.05$, we used the entry for $\alpha = 0.10$ for two-sided tests in that table.) Therefore,

$$\Delta_1 = \delta_1\sqrt{n_1} = \sqrt{1.016}(1.645 + 0.840) = 2.505.$$

We chose $n_1 = n_2 = 50$, which gives $\delta_1 = 2.505/\sqrt{50} = 0.3543$. For the secondary boundary, we chose the PO boundary as discussed before. Then, we calculated the secondary powers for the three methods (the confidence limit method, the conservative method, and the exact method) for selected values of $\Delta_2 = \delta_2\sqrt{n_1}$ for three values of $\rho = 0.3, 0.5, 0.7$. We present the results in Table IV. Note that the power values shown for $\Delta_2 = 0$ are in fact the secondary type I error probabilities.

Inspection of Table IV shows that the confidence limit method achieves a major fraction of the power gains possible compared with the exact method. To give a graphical picture of the relative power advantage of the confidence limit method over the conservative method, we have plotted the percent power gain defined in (10) as a function of Δ_2 for $\rho = 0.3, 0.5, 0.7$ in Figure 5 for the OF1–PO2 boundary combination. We see that the percent power gain increases from about 76% to 84% for $\rho = 0.3$, 74% to 81% for $\rho = 0.5$, and 72% to 79% for $\rho = 0.7$ as Δ_2 increases from 0 to 4 (although the absolute power differences between the three methods tend to 0 as Δ_2 increases). The percent power gain is higher for lower ρ for every Δ_2, which can be easily explained by the fact that the comparisons are carried out with the conservative method, which assumes $\rho = 1$.

7. Extension to two samples

If the trial uses the matched pairs design, then the data can be reduced to the single samples setup in the usual manner by taking the differences between the paired observations on the treatment and the control arms for each patient and applying the methodology given previously. So, we consider two independent samples with n_{ik} patients on arm i in stage k $(i, k = 1, 2)$, where $i = 1$ denotes the treatment arm and $i = 2$ denotes the control (placebo) arm. Let (U_{ijk}, V_{ijk}) denote the observations on the primary and the secondary endpoint on arm i on the jth patient in stage k $(j = 1, \ldots, n_{ik})$. Assume that (U_{ijk}, V_{ijk}) are independent and identically distributed bivariate normal random variates with marginal distributions

Table IV. Secondary power comparisons between the confidence limit method, the conservative method, and the exact method using the OF1–PO2 boundary combination (primary power $= 0.80$, $\Delta_1 = 2.505$, $n_1 = n_2 = 50$, $\alpha = 0.05$).

		Δ_2										
True ρ	Method (ρ)	0	0.4	0.8	1.2	1.6	2.0	2.4	2.8	3.2	3.6	4.0
0.3	$\rho =$ True ρ	0.0500	0.1176	0.2372	0.4049	0.5887	0.7465	0.8552	0.9174	0.9483	0.9619	0.9672
	$\rho = \rho^*$	0.0458	0.1095	0.2243	0.3887	0.5727	0.7339	0.8470	0.9130	0.9462	0.9610	0.9669
	$\rho = 1$	0.0322	0.0819	0.1786	0.3285	0.5102	0.6824	0.8122	0.8930	0.9362	0.9566	0.9652
0.5	$\rho =$ True ρ	0.0485	0.1138	0.2323	0.4046	0.5987	0.7654	0.8753	0.9325	0.9570	0.9659	0.9687
	$\rho = \rho^*$	0.0445	0.1058	0.2194	0.3880	0.5821	0.7527	0.8678	0.9289	0.9555	0.9654	0.9685
	$\rho = 1$	0.0329	0.0823	0.1796	0.3345	0.5263	0.7079	0.8396	0.9146	0.9494	0.9632	0.9679
0.7	$\rho =$ True ρ	0.0445	0.1052	0.2202	0.3972	0.6060	0.7870	0.8993	0.9490	0.9649	0.9687	0.9695
	$\rho = \rho^*$	0.0411	0.0984	0.2087	0.3817	0.5903	0.7753	0.8932	0.9466	0.9642	0.9686	0.9695
	$\rho = 1$	0.0324	0.0804	0.1772	0.3376	0.5435	0.7391	0.8730	0.9385	0.9618	0.9680	0.9694

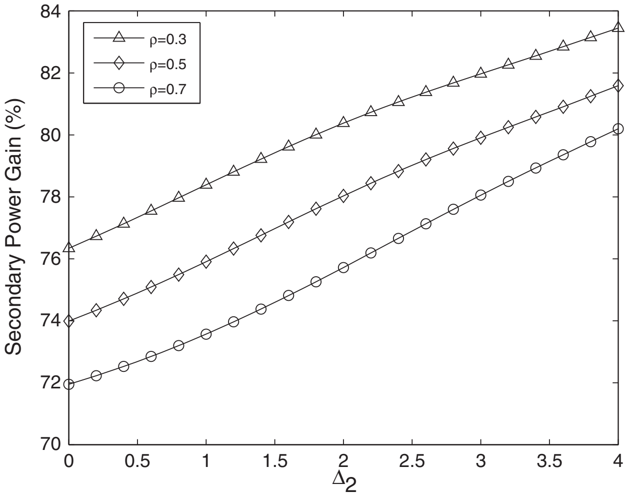

Figure 5. Power gain (%) of the confidence limit method as a function of Δ_2 for $\rho = 0.3, 0.5, 0.7$ and the OF1–PO2 boundary combination

$U_{ijk} \sim N(\xi_i, \sigma_1^2)$, $V_{ijk} \sim N(\eta_i, \sigma_2^2)$, and $\text{corr}(U_{ijk}, V_{ijk}) = \rho \geqslant 0$. Note that this model assumes homoscedasticity, that is, the standard deviations for the two arms are the same (σ_1 for the primary endpoint and σ_2 for the secondary endpoint). All the parameters are unknown except σ_1 and σ_2, which are assumed to be known for convenience (but using their sample estimates is not problematic as seen in Section 5). Let $\delta_1 = (\xi_1 - \xi_2)/\sigma_1$ and $\delta_2 = (\eta_1 - \eta_2)/\sigma_2$ be the standardized treatment effects for the primary and the secondary endpoints, respectively.

To define the test statistic, we first define $\overline{U}_{i\cdot k}$ and $\overline{V}_{i\cdot k}$ as the sample means of the observations U_{ijk} and V_{ijk} averaged over patients $j = 1, \ldots, n_{ik}$, respectively. Then, the first-stage test statistics are given by

$$X_1 = \frac{\overline{U}_{1\cdot 1} - \overline{U}_{2\cdot 1}}{\sigma_1\sqrt{1/n_{11} + 1/n_{21}}} \text{ and } Y_1 = \frac{\overline{V}_{1\cdot 1} - \overline{V}_{2\cdot 1}}{\sigma_2\sqrt{1/n_{11} + 1/n_{21}}}. \tag{11}$$

Next, denote by $\overline{U}_{1\cdot\cdot}$ and $\overline{V}_{1\cdot\cdot}$ the overall sample means of the primary and secondary endpoints data for the treatment arm and $\overline{U}_{2\cdot\cdot}$ and $\overline{V}_{2\cdot\cdot}$ the corresponding overall sample means for the control arm. Also denote by $n_{1\cdot} = n_{11} + n_{12}$ and $n_{2\cdot} = n_{21} + n_{22}$ the total number of patients on the treatment and the control arms, respectively. The cumulative test statistics are given by

$$X_2 = \frac{\overline{U}_{1\cdot\cdot} - \overline{U}_{2\cdot\cdot}}{\sigma_1\sqrt{1/n_{1\cdot} + 1/n_{2\cdot}}} \text{ and } Y_2 = \frac{\overline{V}_{1\cdot\cdot} - \overline{V}_{2\cdot\cdot}}{\sigma_2\sqrt{1/n_{1\cdot} + 1/n_{2\cdot}}}. \tag{12}$$

The correlation structure of (X_1, Y_1, X_2, Y_2) is the same as that given in (5) but with

$$\tau = \sqrt{\left(\frac{n_{1\cdot} + n_{2\cdot}}{n_{1\cdot}n_{2\cdot}}\right)\left(\frac{n_{11}n_{21}}{n_{11} + n_{21}}\right)}. \tag{13}$$

We have assumed that the standard deviations σ_1 and σ_2 are known, but in practice, they must be estimated. The pooled estimates (from the treatment arm and the control arm) of σ_1 and σ_2 at the interim look are given by

$$\widehat{\sigma}_1^{(1)} = \sqrt{\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_{i1}}(U_{ij1} - \overline{U}_{i\cdot 1})^2}{n_{11} + n_{21} - 2}} \text{ and } \widehat{\sigma}_2^{(1)} = \sqrt{\frac{\sum_{i=1}^{2}\sum_{j=1}^{n_{i1}}(V_{ij1} - \overline{V}_{i\cdot 1})^2}{n_{11} + n_{21} - 2}},$$

which are used in (11). Similarly, the overall pooled estimates of σ_1 and σ_2 are given by

$$\widehat{\sigma}_1 = \sqrt{\frac{\sum_{i=1}^{2}\sum_{k=1}^{2}\sum_{j=1}^{n_{ik}}(U_{ijk} - \overline{U}_{i\cdot\cdot})^2}{n_{1\cdot} + n_{2\cdot} - 2}} \text{ and } \widehat{\sigma}_2 = \sqrt{\frac{\sum_{i=1}^{2}\sum_{k=1}^{2}\sum_{j=1}^{n_{ik}}(V_{ijk} - \overline{V}_{i\cdot\cdot})^2}{n_{1\cdot} + n_{2\cdot} - 2}},$$

which are used in (12).

Finally, note that the sample correlation r is computed from the first-stage data consisting of $n_{11}+n_{21}$ pairs (U_{ij1}, V_{ij1}) for $i = 1, 2$; thus, n_1 in formulas (7) and (8) must be replaced by $n_{\cdot 1} = n_{11} + n_{21}$.

8. Example

Burge *et al.* [11] reported the results from the ISOLDE trial, which was a randomized, double-blind, placebo-controlled study of fluticasone propionate (treatment) in patients with moderate to severe chronic obstructive pulmonary disease. Patients were recruited between October 1, 1992 and March 31, 1995 in 18 UK hospitals. After an 8-week run-in period, a total of 751 patients were randomized (376 on the treatment and 375 on the placebo) to receive either 500 μg of the treatment or an identical placebo administered twice daily from a metered dose inhaler. The patients were followed-up for 36 months with visits scheduled every 3 months for spirometry and safety assessments. This was a disease-modifying drug trial where the treatment was intended to slow down the decline of the pulmonary function for patients with chronic obstructive pulmonary disease. The primary outcome measure was the rate of decline in forced expiratory volume at 1 s (FEV1), the trial being anticipated to show that the rate of decline in FEV1 will be smaller in the treatment group than in the placebo group. Forced vital capacity (FVC) was also measured but not used as an efficacy endpoint. There were a total of 612 patients (313 on the treatment and 299 on the placebo) with at least two visits (baseline and final). To account for the dropouts and correlations among the repeated measures, the authors of the study used the mixed model approach.

We will use this dataset but make a few changes in the analysis to fit the setting of the present paper and for the sake of simplicity. First, we will assume that the rate of decline in FEV1 is the primary endpoint and the rate of decline in FVC is the secondary endpoint. Figure 6 shows the rates of decline in FEV1 and FVC in both the treatment and the control groups, which are roughly linear. It is evident that the rates of decline in both the outcome measures are steeper in the control group than in the treatment group. Because of the approximate linear downward trend, we calculated the rate of decline in each case simply by dividing the difference between the final measurement (the timing of which varies from patient to patient depending upon how many visits they completed) and the baseline measurement (at randomization) by the period (in months) between the two measurements. Furthermore, this study employed a fixed sample design with just one final look. We will instead assume a group sequential design with two looks, the interim look being at the quarter point, that is, after the first 153 patients.

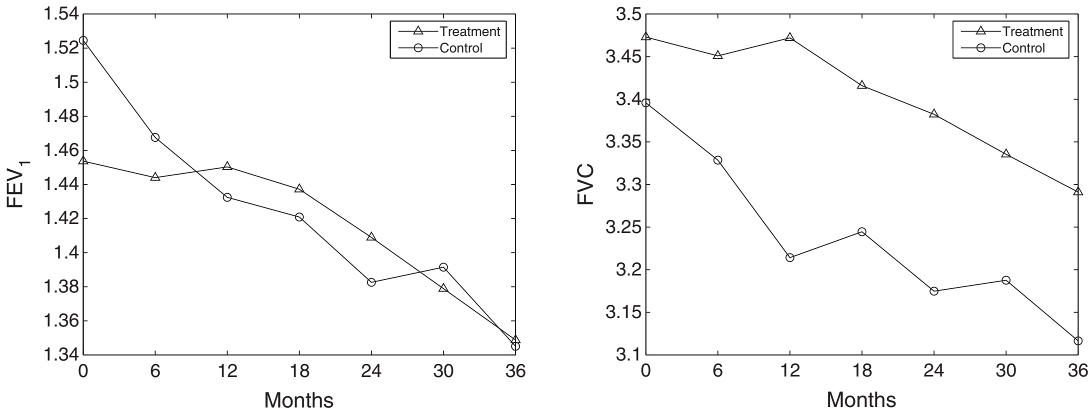

Figure 6. Time series plots of FEV1 (left panel) and FVC (right panel) for the treatment and control groups.

The sample sizes on the treatment and the control arms in the first stage are $n_{11} = 71, n_{21} = 82$. The statistics at the interim look are

$$\overline{U}_{1\cdot 1} = -0.0029, \overline{U}_{2\cdot 1} = -0.0070, \overline{V}_{1\cdot 1} = -0.0089, \overline{V}_{2\cdot 1} = -0.0140,$$

$$\widehat{\sigma}_1^{(1)} = 0.0141, \widehat{\sigma}_2^{(1)} = 0.0301, r = 0.6667.$$

From these summary statistics, we calculate $X_1 = 1.791$ and $Y_1 = 1.040$. We use the OF1–PO2 boundary combination, that is, for FEV1, we use the OF boundary, which depends on the correlation coefficient τ given by (13). Note that this formula involves the observed sample sizes $n_{11}, n_{21}, n_{1\cdot}$, and $n_{2\cdot}$, which are not precisely known at the beginning of the trial ($n_{1\cdot}$ and $n_{2\cdot}$ are not known even at the interim look) because they are the outcomes of randomization. Therefore, we assume equal allocation on the treatment and the control arms. For the interim look at the quarter point, we get $\tau = \sqrt{1/4} = 0.5$ (if we substitute the observed values of $n_{11}, n_{21}, n_{1\cdot}$, and $n_{2\cdot}$, we get $\tau = 0.4988$, which is not too different from 0.5). The corresponding OF1 boundary is $(c_1, c_2) = (2.813, 1.989)$. Because $X_1 < c_1$, we continue sampling to the second stage.

The sample sizes at the second stage are $n_{12} = 242, n_{22} = 217$, which give the total sample sizes on the two arms as $n_{1\cdot} = 313, n_{2\cdot} = 299$. Next, we calculate the PO2 boundary. The sample correlation coefficient $r = 0.6667$ is based on 153 pairs of observations. Using the confidence limit method, we calculate the optimum value of the confidence level $1 - \varepsilon = 0.97$, the upper confidence limit $\rho^* = 0.7540$, and the secondary boundary $(d_1, d_2) = (2.116, 2.116)$. Using the conservative method (assuming $\rho = 1$), we would have $(d_1, d_2) = (2.212, 2.212)$. The cumulative statistics at the final look are

$$\overline{U}_{1\cdot\cdot} = -0.0044, \overline{U}_{2\cdot\cdot} = -0.0080, \overline{V}_{1\cdot\cdot} = -0.0085, \overline{V}_{2\cdot\cdot} = -0.0129, \widehat{\sigma}_1 = 0.0132, \widehat{\sigma}_2 = 0.0280.$$

From these summary statistics, we calculate $X_2 = 3.406$ and $Y_2 = 1.914$. Thus, $X_2 > c_2$, but $Y_2 < d_2$. So, we are able to claim significance on FEV1 but not on FVC.

9. Discussion

This paper has given a powerful method and a table of secondary boundary constants to implement it when the correlation between the primary and the secondary endpoint is unknown and is estimated from the first-stage data in a two-stage group sequential procedure in which the primary endpoint acts as a gatekeeper for the secondary endpoint. The OF1–PO2 boundary combination is shown to give the best secondary power performance among all four boundary combinations; it is already known that the OF1 boundary gives a better primary power performance than the PO1 boundary.

It is natural to ask whether the method can be extended to binary data, which are common in practice. Unfortunately, there are major difficulties as we outline in the following text without getting too much into details. Assume that n_1 independent observations are taken from the treatment and the control group in the first stage and n_2 independent observations are taken from each group in the second stage. Let $p_1 - p_2$ and $q_1 - q_2$ be the differences in success probabilities between the treatment group and the control group for the primary and the secondary endpoints, respectively. The primary and secondary null hypotheses are $H_1 : p_1 - p_2 = 0$ and $H_2 : q_1 - q_2 = 0$ against upper one-sided alternatives. Let $\widehat{p}_1^{(k)} - \widehat{p}_2^{(k)}$ and $\widehat{q}_1^{(k)} - \widehat{q}_2^{(k)}$ be their sample estimates (proportions) based on the cumulative data up to the kth stage ($k = 1, 2$). The first-stage Wald statistics are

$$X_1 = \frac{(\widehat{p}_1^{(1)} - \widehat{p}_2^{(1)})\sqrt{n_1}}{\sqrt{2\widehat{p}^{(1)}(1 - \widehat{p}^{(1)})}}, Y_1 = \frac{(\widehat{q}_1^{(1)} - \widehat{q}_2^{(1)})\sqrt{n_1}}{\sqrt{2\widehat{q}^{(1)}(1 - \widehat{q}^{(1)})}},$$

and the second-stage Wald statistics are

$$X_2 = \frac{(\widehat{p}_1^{(2)} - \widehat{p}_2^{(2)})\sqrt{(n_1 + n_2)}}{\sqrt{2\widehat{p}^{(2)}(1 - \widehat{p}^{(2)})}}, Y_2 = \frac{(\widehat{q}_1^{(2)} - \widehat{q}_2^{(2)})\sqrt{(n_1 + n_2)}}{\sqrt{2\widehat{q}^{(2)}(1 - \widehat{q}^{(2)})}},$$

where $\widehat{p}^{(1)}$ and $\widehat{p}^{(2)}$ are the pooled cumulative estimates of p_1 and p_2 at the first and second stages, respectively, and similarly $\widehat{q}^{(1)}$ and $\widehat{q}^{(2)}$ are the pooled cumulative estimates of q_1 and q_2 at the first and the second stages. For large n_1, n_2, which are typical in clinical trials, the approximate joint normal distribution of (X_1, X_2, Y_1, Y_2) follows from the multivariate central limit theorem. However, the assumption of homoscedasticity, that is, the common variances and the common correlation coefficient, will not be valid unless both the null hypotheses, H_1 and H_2, are true. This is because the covariance matrix between the primary and the secondary endpoint depends on the corresponding success probabilities; on the other hand, for bivariate normal data, the covariance matrix can be specified independently of the means. Unequal variances is not a major problem because they can be estimated precisely with sufficiently large sample sizes. However, unequal correlation coefficients for the treatment and the control groups is problematic because the confidence limit method developed in this paper is applicable only for a common pooled sample correlation coefficient. There is also the question of how well the arctan hyperbolic transformation of the sample correlation coefficient computed from binary data would approximate the normal distribution. Thus, it is debatable whether the method proposed in this article can be readily extended to binary data. This is a topic for future research, which can be addressed via simulation.

Acknowledgements

We thank the referees for their helpful comments and Dr. Stuart Pocock for providing the data used in the ISOLDE trial.

References

1. Tamhane AC, Mehta CR, Liu L. Testing a primary and secondary endpoint in a group sequential design. *Biometrics* 2010; **66**:1174–1184.
2. Hung HMJ, Wang S-J, O'Neill R. Statistical considerations for testing multiple endpoints in group sequential or adaptive clinical trials. *Journal of Biopharmaceutical Statistics* 2007; **17**:1201–1210.
3. Glimm E, Maurer W, Bretz F. Hierarchical testing of multiple endpoints in group-sequential trials. *Statistics in Medicine* 2010; **29**:219–228.
4. Armitage P, McPherson CK, Rowe BC. Repeated significance testing on accumulating data. *Journal of Royal Statistical Society, Series A* 1969; **132**:235–244.
5. Armitage P. *Sequential Medical Trials*. Blackwell: Oxford, 1975.
6. Pocock SJ. Group sequential methods in the design and analysis of clinical trials. *Biometrika* 1977; **64**:191–199.
7. O'Brien PC, Fleming TR. A multiple testing procedure for clinical trials. *Biometrics* 1979; **35**:549–556.
8. Lan KKG, DeMets DL. Discrete sequential boundaries for clinical trials. *Biometrika* 1983; **70**:659–663.
9. Jennison C, Turnbull BW. *Group Sequential Methods with Applications to Clinical Trials*. Chapman & Hall/CRC: Boca Raton, FL, 2000.
10. Berger RL, Boos DD. P-value maximized over a confidence set for the nuisance parameter. *Journal of the American Statistical Association* 1994; **89**:1012–1016.
11. Burger PS, Calverley PMA, Jones PW, Spencer S, Anderson JA, Maslen TK. Randomised, double blind, placebo controlled study of fluticasone propionate in patients with moderate to severe chronic obstructive pulmonary disease: the ISOLDE trial. *British Medical Journal* 2000; **320**:1297–1303.