

A method for utilizing co-primary efficacy outcome measures to screen regimens for activity in two-stage Phase II clinical trials

Michael W Sill^{a,b}, Larry Rubinstein^c, Samuel Litwin^d and Greg Yothers^{e,f}

Background Most Phase II clinical trials utilize a single primary end point to determine the promise of a regimen for future study. However, many disorders manifest themselves in complex ways. For example, migraine headaches can cause pain, auras, photophobia, and emesis. Investigators may believe that a drug is effective at reducing migraine pain and the severity of emesis during an attack. Nevertheless, they could still be interested in proceeding with the development of the drug if it is effective against only one of these symptoms. Such a study would be a candidate for a clinical trial with co-primary end points.

Purpose The purpose of the article is to provide a method for designing a single arm, two-stage clinical trial with dichotomous co-primary end points of efficacy that has the ability to detect activity on either response measure with high probability when the drug is active on one or both measures, while at the same time rejecting the drug with high probability when there is little activity on both dimensions. The design enables early closure for futility and is flexible with regard to attained accrual.

Methods The design is proposed in the context of cancer clinical trials with tumor response and progression-free survival (PFS) status after a certain period. Both end points are assumed to be distributed as binomial random variables, and uninteresting probabilities of success are determined from historical controls. Given the necessity of accrual flexibility, exhaustive searching algorithms to find optimum designs do not seem feasible at this time. Instead, critical values are determined for realized sample sizes using specific procedures. Then accrual windows are found to achieve a design's desired level of significance, probability of early termination (PET), and power.

Results The design is illustrated with a clinical trial that examined bevacizumab in patients with recurrent endometrial cancer. This study was negative by tumor response but positive by 6-month PFS. The procedure was compared to modified procedures in the literature, indicating that the method is competitive.

Limitations Although the procedure allows investigators to construct designs with desired levels of significance and power, the PET under the null hypothesis is smaller than for single end point studies.

Conclusions The impact of adding an additional end point on the sample size is often minimal, but the study gains sensitivity to activity on another dimension of treatment response. The operating characteristics are fairly robust to the level of association between the two end points. Software is available online. *Clinical Trials* 2012; 9: 385–395. <http://ctj.sagepub.com>

^aGynecologic Oncology Group Statistical and Data Center, Roswell Park Cancer Institute, Buffalo, NY, USA, ^bState University of New York at Buffalo, Buffalo, NY, USA, ^cBiometric Research Branch, National Cancer Institute, Bethesda, MD, USA, ^dBiostatistics Facility, Fox Chase Cancer Center, Philadelphia, PA, USA, ^eDepartment of Biostatistics, University of Pittsburgh, Pittsburgh, PA, USA, ^fNational Surgical Adjuvant Breast and Bowel Project Biostatistical Center, University of Pittsburgh, Pittsburgh, PA, USA

Author for correspondence: Michael W Sill, Gynecologic Oncology Group Statistical and Data Center, Roswell Park Cancer Institute, Elm and Carlton Streets, Buffalo, NY 14263, USA.

Email: msill@gogstats.org

© The Author(s), 2012

Reprints and permission: <http://www.sagepub.com/journalsPermissions.nav>

Downloaded from ctj.sagepub.com at NIH LIBRARY on October 27, 2015

10.1177/1740774512450101

Introduction

Phase II studies evolved over time to simultaneously manage several goals of a clinical trial while maintaining their modest sample size. Initially, they were simple studies, designed to distinguish between two probabilities of response with α level of significance and power $(1-\beta)$. Gehan [1] proposed a design with a futility rule in 1961, which rejected a drug early in a clinical trial if there were no observed responses in the first stage. A more general design that allows multistage testing for arbitrary values of π_{r0} and π_{r1} was provided by Fleming [2]. Simon [3] then proposed a two-stage design that had either optimal or minimax properties in 1989. His solution was extended with flexible designs that allowed for deviations from the targeted sample sizes in 1998 [4,5].

In addition to incorporating flexible, interim futility analyses, several authors proposed utilizing more than one primary end point. Bryant and Day [6], Thall and Cheng [7], and Conaway and Petroni [8] use the number of patients who have severe adverse events in their decision rules for recommending further study.

Some authors refined response to therapy into three ordered classes. In the arena of oncology, patient responses can be classified as progressive disease, stable disease, and tumor response so that therapies capable of reducing the proportion with progressive disease or increasing the proportion with tumor responses are of interest [9]. Other authors differentiate complete tumor response from partial responses and study the impact of employing both of these positive outcomes to trial characteristics [10–12].

Another approach examined in the literature is the utilization of two (or more) fundamentally different measures of treatment efficacy [13,14]. Examples of co-primary efficacy end points include the severity of angina pectoris and shortness of breath for studies involving coronary artery disease. Unlike the gradation of a univariate response into separate categories, these designs consider the multivariate nature of response to treatment that is complicated by variable associations and study objectives. Bayesian approaches have also been proposed with applications in jointly evaluating efficacy and toxicity as well as multiple efficacy outcomes [15,16]. The current article provides investigators with a method for obtaining a two-stage trial design with an interim futility rule where interest is focused on detecting activity on either of two primary response variables. The design is also flexible with regard to the attained sample size. That is to say, the design does not require a precise sample size for each stage. Instead, accrual windows (or allowable accrual ranges) like those seen in Chen

and Ng [4] are provided along with design characteristics such as average power or minimum power (over the accrual range).

Methodology

Our methodology will be developed in the context of Phase II cancer clinical trials. In this setting, many drugs have been evaluated with tumor response as defined by Therasse *et al.* [17]. The probability of response will be designated with π_r . Another variable of interest in Phase II oncology is the probability that a patient survives without experiencing a progression of disease for a specified period of time. We will use 6 months as a matter of convenience and denote this binomial variable as '6-month progression-free survival (PFS)'. The probability of this event will be designated with π_s .

The null hypothesis is formulated as follows: $H_0 : \pi_r \leq \pi_{r0}$ and $\pi_s \leq \pi_{s0}$, where π_{r0} and π_{s0} are specified values (obtained from historical data) that are believed to be uninteresting or comparable to the current standard of care. The alternative hypothesis is formulated as follows: $H_1 : \pi_r \geq \pi_{r1} = \pi_{r0} + \Delta_r$ or $\pi_s \geq \pi_{s1} = \pi_{s0} + \Delta_s$, where Δ_r and Δ_s are the (minimal) clinically significant improvements in the proportion responding and with 6-month PFS, respectively.

Relationships between critical values and the design's operating characteristics

The joint distribution of the number of patients who respond or have 6-month PFS is provided along with the parameters in Table 1. Note that $n_{(k)}$ is the sample size for stage $k = 1, 2$ of the trial, $X_{r(k)}$ is the number of patients who have an objective response in stage k , and $X_{s(k)}$ is the number of patients with 6-month PFS. It can be shown that the joint distribution of $X_{ij(k)}$, where $i = 1, 2$ and $j = 1, 2$, is multinomial with the corresponding parameters listed in Table 1 under the restrictions $\pi_{22} = 1 - \pi_{11} - \pi_{12} - \pi_{21}$ and $X_{22(k)} = n_{(k)} - X_{11(k)} - X_{12(k)} - X_{21(k)}$. The probability mass function (PMF) of this distribution is

$$g(x_{ij(k)}; i = 1, 2; j = 1, 2; k = 1, 2) = \frac{n_{(k)}!}{x_{11(k)}! x_{12(k)}! x_{21(k)}! x_{22(k)}!} \pi_{11}^{x_{11(k)}} \pi_{12}^{x_{12(k)}} \pi_{21}^{x_{21(k)}} \pi_{22}^{x_{22(k)}} \quad (1)$$

The drug is rejected at Stage 1 if $X_{r(1)} \leq C_{r(1)}$ and $X_{s(1)} \leq C_{s(1)}$ or at Stage 2 if $X_r \leq C_r$ and $X_s \leq C_s$ where $C_{r(1)}$ and $C_{s(1)}$ are the Stage 1 critical values for the number of patients who have a response and the number with 6-month PFS, respectively; $X_r = X_{r(1)} + X_{r(2)}$, $X_s = X_{s(1)} + X_{s(2)}$, and C_r and C_s are the critical

Table 1. Parameters for the associated probabilities of tumor response and 6-month PFS are specified on the left side. For example, the joint probability of having a response and 6-month PFS is given by π_{11} , and the marginal probability of having a response is given by $\pi_r = \pi_{11} + \pi_{12}$. The number of people in the trial at stage k with these qualities is provided on the right side. For example, the number of people in stage k with tumor response and have 6-month PFS is provided by $X_{11(k)}$.

		Table for parameters		Table for data		
		PFS > 6 months		PFS > 6 months		
		Yes	No	Yes	No	
Response	Yes	π_{11}	π_{12}	π_r	$X_{11(k)}$	$X_{12(k)}$
	No	π_{21}	π_{22}	π_{2+}	$X_{21(k)}$	$X_{22(k)}$
		π_s	π_{+2}	$X_{s(k)}$	$X_{+2(k)}$	$X_{r(k)}$
						$X_{2+(k)}$

PFS: progression-free survival.

values at the end of Stage 2. The following relationships hold in general

$$X_{12(k)} = X_{r(k)} - X_{11(k)} \tag{2}$$

$$X_{21(k)} = X_{s(k)} - X_{11(k)} \tag{3}$$

$$X_{22(k)} = n_{(k)} - X_{11(k)} - X_{12(k)} - X_{21(k)} \tag{4}$$

To determine the probability of rejecting the drug after a particular stage using equation (1), it is helpful to define a PMF and a cumulative distribution function (CDF) in terms of $n_{(k)}, X_{s(k)}$, and $X_{r(k)}$.

$$P(X_{r(k)} = x_{r(k)}, X_{s(k)} = x_{s(k)}) = \sum_{x_{11(k)} = \max\{0, x_{r(k)} + x_{s(k)} - n_{(k)}\}}^{\min\{x_{r(k)}, x_{s(k)}\}} g(x_{ij(k)}) \tag{5}$$

$$f(n_{(k)}, x_{r(k)}, x_{s(k)}) = \sum_{x_{11(k)} = \max\{0, x_{r(k)} + x_{s(k)} - n_{(k)}\}}^{\min\{x_{r(k)}, x_{s(k)}\}} g(x_{ij(k)})$$

$$P(X_{r(k)} \leq r, X_{s(k)} \leq s) = F(n_{(k)}, r, s) = \sum_{x_{r(k)}=0}^r \sum_{x_{s(k)}=0}^s \sum_{x_{11(k)} = \max\{0, x_{r(k)} + x_{s(k)} - n_{(k)}\}}^{\min\{x_{r(k)}, x_{s(k)}\}} g(x_{ij(k)}) \tag{6}$$

The probability of early termination (PET), which is the probability of rejecting the drug after the first stage, can be calculated simply with the CDF and by using the first-stage parameters, that is

$$PET = F(n_{(1)}, C_{r(1)}, C_{s(1)}) \tag{7}$$

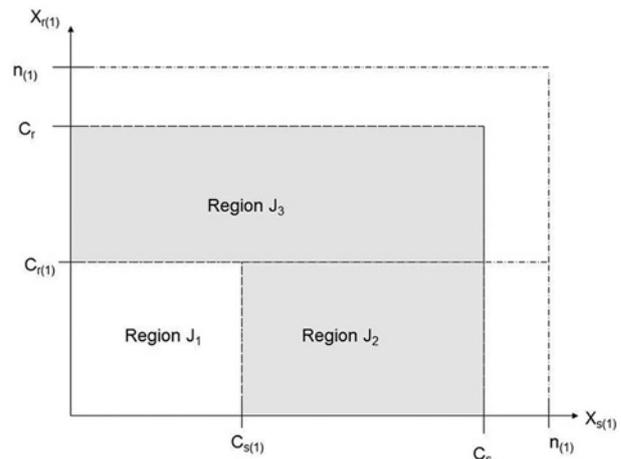


Figure 1. Sample space for $X_{r(1)}$ and $X_{s(1)}$ divided into several important regions. Region J_1 is the set of outcomes that lead the trial into early termination. Regions J_2 and J_3 are areas that allow the trial to proceed to Stage 2 but do not necessarily lead to the rejection of the null hypothesis. The complement of $J_1 \cap J_2 \cap J_3$ would ultimately lead to rejection of the null hypothesis.

where $n_{(1)}$ is the first-stage sample size. In order for the drug to be rejected after the second stage, it is required that the outcome after the first stage does not lie within the drug's rejection region (i.e., $X_{r(1)} \leq C_{r(1)}$ and $X_{s(1)} \leq C_{s(1)}$) but that the outcome in the second stage lies within the drug's rejection region (i.e., $X_r \leq C_r$ and $X_s \leq C_s$). In order for this condition to be true, it is required that the following condition hold: $X_{r(1)} > C_{r(1)}$ or $X_{s(1)} > C_{s(1)}$ and, simultaneously, that $X_{r(1)} \leq C_r$ and $X_{s(1)} \leq C_s$ for the first-stage outcome. Using Figure 1, this region corresponds to the union of regions J_2 and J_3 (note that it is possible for C_s and C_r to be greater than $n_{(1)}$).

To calculate the probability that the drug is rejected in the second stage, it is important to note that $X_r - X_{r(1)}$ and $X_s - X_{s(1)}$ are marginal totals equal to $X_{r(2)}$ and $X_{s(2)}$ whose cells have a multinomial distribution with the parameters listed in Table 1 with a sample size of $n_{(2)}$. $X_s \leq C_s$ if and only if $X_s - X_{s(1)} \leq C_s - X_{s(1)}$, and $X_r \leq C_r$ if and only if $X_r - X_{r(1)} \leq C_r - X_{r(1)}$. It follows that the probability of rejecting the drug in Stage 2 is

$$P(\text{Rejecting drug in Stage 2} \cap \text{Trial passed Stage 1}) = \sum_{(X_{s(1)}, X_{r(1)}) \in J_2 \cup J_3} f(n_{(1)}, X_{r(1)}, X_{s(1)}) F(n - n_{(1)}, C_r - X_{r(1)}, C_s - X_{s(1)}) \tag{8}$$

where $J_2 \cup J_3$ is Region J_2 union Region J_3 and $n = n_{(1)} + n_{(2)}$. The total probability of rejecting the drug is the sum of the probabilities of rejecting the drug in each stage. That is, $P(\text{Rejecting drug}) = PET +$

$P(\text{Rejecting drug in Stage 2})$, and the power of the study is simply the probability of accepting the drug or alternatively rejecting H_0 .

$$\text{Power} = P(\text{Reject } H_0) = 1 - [PET + P(\text{Reject drug in Stage 2})] \quad (9)$$

Search for designs with desirable operating characteristics

Important features of the design include a high PET under the null hypothesis, low power under the null hypothesis, and high power under the alternatives. This design is useful for trials where the investigators want to detect true activity on either end point with high probability. There are three design parameters that are believed to be of particular interest

$$\begin{aligned} \alpha &= P(\text{reject } H_0 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0}) \\ \beta_r &= P(\text{reject drug} \mid \pi_r = \pi_{r0} + \Delta_r, \pi_s = \pi_{s0}) \\ \beta_s &= P(\text{reject drug} \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0} + \Delta_s) \end{aligned}$$

where α is the probability of a type I error, β_r is the probability of a type II error when the agent is clinically active by response but is not capable of stabilizing the disease for long periods, and β_s is the probability of a type II error when the regimen is clinically active by stabilization of disease but does not significantly reduce tumor burden. For a particular set of critical boundaries, $C_{s(1)}, C_{r(1)}, C_s$, and C_r , the following quantities can be found

$$\begin{aligned} PET_{H0} &= P(\text{reject drug Stage 1} \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0}) \\ PET_{Hr} &= P(\text{reject drug Stage 1} \mid \pi_r = \pi_{r0} + \Delta_r, \pi_s = \pi_{s0}) \\ PET_{Hs} &= P(\text{reject drug Stage 1} \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0} + \Delta_s) \\ TPRT_{H0} &= P(\text{reject drug} \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0}) = 1 - \alpha \\ TPRT_{Hr} &= P(\text{reject drug} \mid \pi_r = \pi_{r0} + \Delta_r, \pi_s = \pi_{s0}) = \beta_r \\ TPRT_{Hs} &= P(\text{reject drug} \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0} + \Delta_s) = \beta_s \end{aligned}$$

where ‘TPT’ stands for the total probability of rejecting the treatment.

Optimal and minimax designs

Exhaustive searching algorithms for optimal or minimax designs have not been developed for co-primary end points yet. Conceptually, they can be developed first by finding the set of designs that meet the specified requirements of the investigator such as $\alpha \leq 0.10$, $\beta_r \leq 0.10$, and $\beta_s \leq 0.10$. To find the optimal design among all designs meeting the

investigator’s specifications, the design associated with the minimal value of $E[N_t]$ under the null hypothesis would be selected where

$$\begin{aligned} E[N_t] &= PET_{H0} \cdot n_{(1)} + (1 - PET_{H0})(n_{(1)} + n_{(2)}) \quad (10) \\ E[N_t] &= n_{(1)} + (1 - PET_{H0})n_{(2)} \end{aligned}$$

where N_t is the total sample size, which is a random variable equal to $n_{(1)}$ with probability PET_{H0} and $n = n_{(1)} + n_{(2)}$ otherwise. The minimax design would be a design (with specifications) associated with the smallest n , called $\min\{n\}$. If several designs existed with $n = \min\{n\}$, then the one that minimizes equation (10) would be utilized.

Flexible optimal and flexible minimax designs

Flexible optimal designs, developed in a manner similar to Chen and Ng [4], would provide designs where accrual windows are allowed such as $n_L \leq N_{(1)} \leq n_U$ and $N_L \leq N \leq N_U$, where $N_{(1)}$ and N are the first-stage and total sample sizes, respectively. $N_{(1)}$ and N are random variables with realized values equal to $n_{(1)}$ and n , respectively. The size of these windows is usually eight patients wide. Assuming a uniform accrual distribution (e.g., $P(N_{(1)} = n_{(1)}) = 1/8$ for $n_L \leq n_{(1)} \leq n_U$ and $P(N_{(1)} = n_{(1)}, N=n \mid \text{Trial passed Stage 1}) = 1/64$), these designs can be characterized by mean values of α , β_r , and β_s over all possible accrual combinations. Alternatively, the design can be characterized by more conservative values of $\max\{\alpha\}$, $\max\{\beta_r\}$, and $\max\{\beta_s\}$. Then, a searching algorithm can be developed in a manner similar to the ‘rigid’ designs as discussed earlier. The implementation of these ideas is challenging because the total number of designs is large.

A Green–Dahlberg searching algorithm

Green and Dahlberg [5] proposed a simple, flexible design for clinical trials that utilize a single, dichotomous primary end point. Their design tests a one-sided alternative hypothesis against the null hypothesis (stating the drug is inactive) after the first stage. If they are able to reject the alternative hypothesis, they declare the drug to be unworthy at the interim analysis. Otherwise, they continue to the second stage.

We propose a similar method for the interim futility. Specifically, we propose using the critical values $C_{r(1)}$ and $C_{s(1)}$ that maximize PET_{H0} from among all designs that limit $PET_{Hr} \leq \beta_r/2$ and $PET_{Hs} \leq \beta_s/2$ for a particular, realized first-stage sample size $n_{(1)}$. For purposes of computational speed, these values are calculated under the assumption that $X_{r(1)}$ is independent of $X_{s(1)}$.

If the design proceeds to the second stage, an eventual sample size of n patients will be realized. We propose several methods (with available software) for determining the critical values at the final stage, C_r and C_s . The first method utilized in clinical trials (called the minimum C method) found C_r and C_s such that the following cost function was minimized

$$C = (1 - TPRT_{H0})^2 + (TPRT_{Hr})^2 + (TPRT_{Hs})^2 \quad (11)$$

These decision rules tend to yield satisfactory designs if $\alpha \approx \beta_r \approx \beta_s$ is desired. Some researchers may prefer designs with $\alpha \leq 0.05$ and $\beta_r \approx \beta_s \approx 0.20$. To help achieve designs with characteristics such as these, an 'alpha restricted method' is offered. This method searches for a C_r and C_s that minimizes the quantity, $\max\{TPRT_{Hr}, TPRT_{Hs}\}$, among all designs where $TPRT_{H0} \geq 1 - \alpha$, assuming that $X_{n(i)}$ is independent of $X_{s(i)}$, $i = 1, 2$.

This decision rule may not yield a design with precisely the desired operating characteristics for the realized sample sizes obtained in the clinical trial, $n_{(1)}$ and n , but a design that follows these procedures will yield unambiguous decision criteria. When $n_{(1)}$ and n deviate from the targeted values, say $n_{(1)}^*$ and n^* , then the realized operating characteristics should not be substantially different from the planned study characteristics as long as the deviations are relatively small.

Searching for flexible designs that obtain the planned α , β_r , and β_s

The first step in obtaining a flexible design is the tabulation of $C_{r(1)}$, $C_{s(1)}$, C_r and C_s along with the design's operating characteristics for all values of n such that $25 \leq n \leq 100$ (or some other suitable range) and $15 \leq n_{(1)} \leq n - \min\{n_{(2)}\}$ (where $\min\{n_{(2)}\}$ is the minimum Stage 2 sample size).

The next step is to characterize the flexible designs by the minimum values of $N_{(1)}$ and N (i.e., by n_L and N_L), where $n_L \leq N_{(1)} \leq n_U$, $N_L \leq N \leq N_U$, $n_U = n_L + (w_1 - 1)$, $N_U = N_L + (w_2 - 1)$, and w_1 and w_2 are the allowable accrual range (typically $w_1 = w_2 = 8$). The characterization can be done with mean values of PET_{H0} , $TPRT_{H0}$, $TPRT_{Hr}$ and $TPRT_{Hs}$. Alternatively, they can be characterized by the $\min\{PET_{H0}\}$, $\min\{TPRT_{H0}\}$, $\max\{TPRT_{Hr}\}$, and $\max\{TPRT_{Hs}\}$.

The subsequent design criterion is to require the average PET_{H0} or the $\min\{PET_{H0}\}$ to be at least some minimal value and find the smallest n_L where this criterion is satisfied. Given a fixed first-stage sample size, the final step is to find the smallest value of N_L so that the average (or maximum values) of $TPRT_{Hr}$ and $TPRT_{Hs}$ is $\leq \beta_r$ and β_s , respectively.

Once a design is found under the assumption of independence, it is characterized under different degrees of association. Investigation has shown that PET_{H0} tends to be moderately dependent on the level of association between the two end points, whereas $TPRT_{Hr}$ and $TPRT_{Hs}$ tend to be fairly robust against this nuisance parameter [6,13]. Also, it can be shown that both PET_{H0} and $TPRT_{H0}$ are higher for the positive association of the two end points (the anticipated relationship) than for independence since, for positively correlated co-primary end points, the probability that both will be discouraging is greater than the product of the individual probabilities that each will be discouraging. This is an important characteristic since it assures that assuming independence of the two end points is conservative in the sense that this assumption yields an upper bound on type 1 error and a lower bound on PET_{H0} , assuming that violations of this assumption are always in the direction of positive association of the two end points. Programs and instructions can be downloaded from <http://www.gog.org/sdcstaff/MikeSill/> and clicking on the link for 'co-primary Phase II studies'.

Illustration of study design for GOG 0229E

Study design

The first study that utilized the proposed methods was a protocol called GOG 0229E, which investigated the effects of bevacizumab on patients with recurrent or persistent endometrial cancer. The values for design parameters of GOG 0229E (π_{r0} and π_{s0}) were obtained from the results of a series of prior protocols in GOG 0129 and GOG 0229 (see Table 1 in the published results of this study by Aghajanian *et al.* [18] for further details). The patients enrolled into GOG 0229E were expected to behave similar to those eligible in the historical studies if the agent was not clinically active.

The null hypothesis was formulated as follows: $H_0 : \pi_r \leq 0.10$ and $\pi_s \leq 0.15$. With $\Delta_r = 0.20$ and $\Delta_s = 0.20$ considered clinically significant, the alternative region of interest is specified with $H_1 : \pi_r \geq 0.30$ or $\pi_s \geq 0.35$. A design was found using the 'Minimum C method' along with using average values of PET_{H0} , $TPRT_{H0}$, $TPRT_{Hr}$ and $TPRT_{Hs}$. The accrual window for GOG 0229E was only five patients wide (in contrast to the current designs of eight patients). The targeted accrual for the first stage was set to 19 eligible and evaluable patients. The cumulative targeted accrual for the second stage was set to 42 patients. The critical values for each stage are provided in Table 2.

The operating characteristics of these designs are provided in the following using the usual definition

Table 2. Table of critical values for Stage 1 and Stage 2. The critical values depend on the number of people recruited at each stage.

$n_{(1)}$	Stage 1	Stage 2				
	$(C_{r(1)}, C_{s(1)})$	$(C_r, C_s) n$				
		40	41	42	43	44
17	(2,2)	(7,9)	(7,9)	(7,10)	(7,10)	(8,10)
18	(2,2)	(7,9)	(7,9)	(7,10)	(8,10)	(8,10)
19	(2,3)	(7,9)	(7,9)	(7,10)	(8,10)	(8,10)
20	(2,3)	(7,9)	(7,9)	(7,10)	(8,10)	(8,10)
21	(2,3)	(7,9)	(7,9)	(7,10)	(8,10)	(8,10)

Table 3. Average power and PET of the study when the variables are independent (e.g., $\pi_{11} = \pi_r \cdot \pi_s$), in the top set, and when they are partially dependent (e.g., $\pi_{11} = 0.90\min\{\pi_r, \pi_s\}$), and fully dependent, $\pi_{11} = \min\{\pi_r, \pi_s\}$, in the bottom set

Value of π_{11}	π_r	π_s	Power (%)	PET (%)
$\pi_r \cdot \pi_s$	0.10	0.15	9.0	41.3
	0.30	0.15	93.4	2.8
	0.10	0.35	92.2	2.7
$0.90\min\{\pi_r, \pi_s\}$	0.30	0.35	99.5	0.2
	0.10	0.15	7.2	52.5
	0.30	0.15	91.8	4.7
$\min\{\pi_r, \pi_s\}$	0.10	0.35	91.4	3.7
	0.30	0.35	96.1	1.8
	0.10	0.15	6.8	54.5
	0.30	0.15	91.7	4.9
	0.10	0.35	91.4	3.8
	0.30	0.35	94.7	2.6

PET: probability of early termination.

of power in three relatively extreme cases: (1) when response is independent of 6-month PFS and (2) when there is a relatively high (or full) association between these outcomes. The actual association is believed to be within this range. To assess the operating characteristics when the two primary end points are not independent, the probability calculations were done with the assumption that the joint probability was $\pi_{11} = 0.90 \cdot \min\{\pi_r, \pi_s\}$ or $\min\{\pi_r, \pi_s\}$, which carries a fairly high degree of association. As can be seen from this example in Table 3, the power of the study is not highly dependent on the level of association between response and PFS at 6 months.

Study results

The results of the study are published by Aghajanian *et al.* and reproduced here for illustrative purposes [18,19]. The study had an unusually high accrual rate for this population with many institutions enrolling patients in the last week before study closure. The first stage of the study enrolled 23 patients

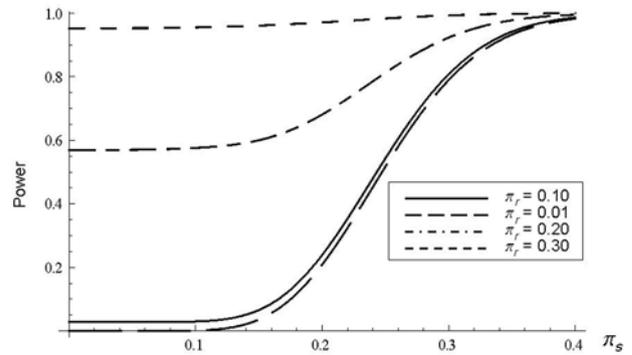


Figure 2. Power curves for π_s when the true response rates are 1%, 10%, 20%, and 30% for the realized sample size for GOG 0229E, which had sample sizes of 21 and 31 patients in Stages 1 and 2, respectively. The calculated power was done under the assumption of independence.

with 2 patients being excluded for not meeting eligibility criteria leaving 21 eligible and evaluable patients. Of these, one patient responded and five had 6-month PFS. Based on Table 2 ($C_{r(1)} = 2$ and $C_{s(1)} = 3$) with evidence of good tolerability, a decision was made to open the study to the second stage.

Stage 2 accrued 33 additional patients with 2 exclusions based on eligibility criteria, leaving a cumulative accrual of 52 patients. This sample size fell outside the targeted window, so a specific rejection boundary had to be calculated for this particular accrual. Using the methodology listed earlier with the first-stage accrual, the second-stage critical boundary was easily adjusted to reflect the larger cohort ($C_r = 9$ and $C_s = 12$). The larger sample size provided more information to reduce the error probabilities (e.g., $\alpha = 0.066$, $\beta_r = 0.039$, and $\beta_s = 0.058$ under independence, and $\alpha = 0.053$, $\beta_r = 0.047$, and $\beta_s = 0.066$ under high association). Figure 2 provides power curves of the parameter π_s for the realized sample size at various values of π_r . When π_r is small and in the null parameter space, power is small when $\pi_s \leq 0.15$ but increases to about 95% when $\pi_s = 0.35$. When $0.10 < \pi_r = 0.20 < 0.30$, the response rate is between the null value and the minimal threshold, so power is about 60% when $\pi_s \leq 0.15$ but increases to 95% as $\pi_s \rightarrow 0.35$. Finally, when $\pi_r = 0.30$, which is in the alternative space, power is high regardless of the value of π_s . Figure 3 provides a contour plot of the power function.

The observed number of patients with responses or who had 6-month PFS was 7 (13.5%) and 21 (40.4%), respectively. Since $X_s = 21 > 12 = C_s$, the agent was deemed active and warranted further investigation [18,19]. The regimen’s response rate was close to the null value of 10% and was insufficient (on its own merits) to open the study to

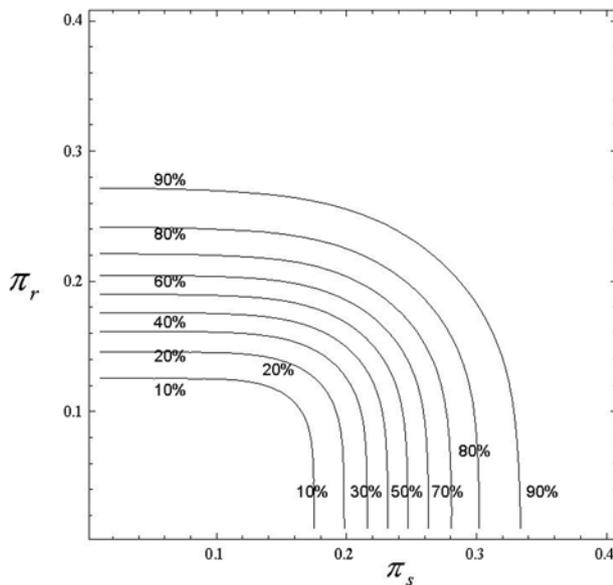


Figure 3. A contour plot of the power function on π_s and π_r for the realized sample size assuming independence.

the second stage or declare it worthy of further investigation.

Comparisons with other procedures

Alternative procedures

To compare the methods presented in this article to simple modifications of existing procedures, an adjusted Simon's procedure was examined. Simon's procedure can be adjusted for a co-primary design by providing the algorithm 1/2 the intended probability of a type I error (α). For example, if the desired overall probability of a type I error in a co-primary design is 10%, then the user would enter 0.05 for α into the program. This method approximately limits the overall study's statistical size of the co-primary procedure to approximately α . It should be noted, however, that the study's size is not guaranteed to be strictly less than α as suggested by a Bonferroni correction since an interim analysis using two variables for proceeding onto Stage 2 can substantially increase the probability of proceeding to Stage 2. Yet, the rule works fairly well in practice. The intended probability of a type II error (β) was not adjusted since the investigator is expected to be interested in detecting activity on either scale with the desired level of power ($1 - \beta$). When the null probabilities for both scales are equal to the same levels of clinical significance (i.e., $\Delta_r = \Delta_s$), Simon's modified procedure can be used to obtain the required sample sizes (Stages 1 and 2) and the

rejection boundaries for each measure of interest. It is important to remember that typical output characterizing Simon's method cannot be used to characterize a co-primary design. For example, the PET for Simon's design is listed as 71.7% when $\pi_r = 0.05$, $\alpha = 0.05$, $\beta = 0.10$, and $\Delta_r = 0.15$. Yet, under independence, the PET is reduced to $0.717^2 = 0.514$ (see the first row of Table 4). Instead, programs utilizing the joint distribution are needed to calculate the operating characteristics.

For cases where the null probabilities are not the same or the interval for clinical significance is different (i.e., $\Delta_r \neq \Delta_s$), then a two-step procedure using univariate methods can be used. For example, Simon's procedure was used to examine the sample size requirements on both scales, and the scale requiring the larger total sample size was used to determine the study's interim and final sample size as well as the rejection boundaries for this more demanding parameter. However, it was not appropriate to use the same rejection boundaries for both measures of efficacy. In this case, another procedure by Schultz was used to find the rejection boundaries for the other scale (e.g., response) conditioned on the interim and final sample sizes [20]. Again, a value of 1/2 the intended α was provided to the algorithm without modification to β . This procedure is referred to as the modified Simon-Schultz method.

Simon's optimal procedure determined the sample size for the two rows per parameter setting in the tables. The procedures using the methods in this article at Simon's sample sizes are labeled as 'flexible co-primary-binomial method' (Flx-CoE). Five features were inspected under the assumption of independent clinical outcomes: PET, the expected sample size under the null hypothesis ($E[N_t | H_0]$), the realized size of the test, and the realized statistical power under two alternative hypotheses (H_r and H_s).

Finally, a so-called optimal procedure was examined (labeled 'optimal co-primary-binomial' (Opt-CoE)), which utilized the procedures of this article to find the smallest $E[N_t | H_0]$ subject to the desired statistical size and power. These designs were characterized by the same five features as described earlier.

Results

For the designs where $\Delta_r = \Delta_s = 0.15$, $\alpha = 0.10$ and $\beta = 0.10$, Simon's adjusted procedure was quite competitive with the Flx-CoE procedure when $\pi_{r0} = \pi_{s0}$. Often both procedures yielded identical rejection boundaries. For the case where $\pi_{r0} = \pi_{s0} = 10\%$, the size of Simon's procedure was slightly greater than 10%, so its power was slightly greater than the Flx-CoE procedure especially under H_s . All the procedures

Table 4. Operating characteristics for three methods: Simon's adjusted design (or a modified Simon–Schultz procedure), a flexible co-primary-binomial method evaluated at Simon's sample sizes (Flx-CoE), and an optimal co-primary-binomial method (Opt-CoE)

Method	π_{r0} (%)	π_{s0} (%)	n_1	n	PET (%)	$E[N_t H_0]$	Size (%)	Power H_r (%)	Power H_s (%)
Simon ^a	5	5	21	41	51	30.7	9.3	91.4	91.4
Flx-CoE					51	30.7	9.3	91.4	91.4
Opt-CoE			21	40	51	30.2	8.6	90.6	90.6
Simon ^a	10	10	21	66	42	47.1	10.4	92.5	92.5
Flx-CoE					42	47.1	7.9	92.4	89.8
Opt-CoE			27	57	52	41.5	9.5	90.7	90.7
Simon ^a	20	20	37	83	47	61.4	9.9	91.5	91.5
Flx-CoE					47	61.4	9.9	91.5	91.5
Opt-CoE			37	78	47	58.7	9.2	90.0	90.0
Simon ^a	30	30	39	100	38	76.7	9.6	91.9	91.9
Flx-CoE					38	76.7	9.6	91.9	91.9
Opt-CoE			49	91	52	69.4	9.4	90.0	90.0
Simon ^a	70	70	25	79	43	55.5	10.3	92.5	92.5
Flx-CoE					43	55.5	8.3	92.3	89.9
Opt-CoE			34	68	54	49.8	9.8	90.6	90.6
Simon ^a	80	80	19	42	58	28.6	9.7	91.6	91.6
Flx-CoE					42	32.4	7.2	85.2	92.2
Opt-CoE			18	43	53	29.7	8.8	91.7	91.7
Simon ^{b,c}	5	10	21	66	22	56.1	10.0	98.2	94.3
Flx-CoE					26	54.2	9.9	96.8	95.3
Opt-CoE			28	46	58	35.5	9.9	90.8	90.1
Simon ^b	10	15	30	82	13	75.2	8.7	96.9	95.0
Flx-CoE					46	58.1	9.5	93.0	93.6
Opt-CoE			37	61	58	47.2	9.3	90.6	90.3
Simon ^b	10	25	37	99	7	94.6	9.5	98.8	94.9
Flx-CoE					48	69.4	9.3	96.3	93.1
Opt-CoE			38	74	54	54.5	9.5	90.2	90.4

PET: probability of early termination.

Note that $\Delta_r = \Delta_s = 0.15, \alpha = \beta = 0.10$.

^aProcedure utilized the modified Simon procedure as defined earlier.

^bProcedure utilized the modified Simon–Schultz method.

^cThe Simon–Schultz method recommended always proceeding to the second stage. In this case, we required at least one response before proceeding to the second stage.

controlled the probability of a type I error fairly closely. Simon's procedure slightly exceeded the targeted power of 90% in many cases, which may be considered an attractive feature since power drops slightly when response is positively correlated with the PFS outcome. The modified Simon–Schultz method performed poorly on several points. It had a low PET and power higher than the desired level (90%). This made the expected sample size under the null hypothesis considerably larger than the co-primary procedures.

When examining the expected sample size under the null hypothesis, the Opt-CoE procedure seemed to perform the best. The only exception was when $\pi_{r0} = 80\%$ and $\pi_{s0} = 80\%$, but the expected sample sizes were quite close in this case.

For the designs where $\Delta_r = \Delta_s = 0.20, \alpha = 0.05$, and $\beta = 0.20$, Table 5 shows that Simon's procedure

often equaled or outperformed the Flx-CoE method by the expected sample sizes. Some of these comparisons may be considered 'unfair' since the actual size of Simon's procedure occasionally was greater than 5%. Although Simon often beats the Flx-CoE method by PET and $E[N_t | H_0]$, the Flx-CoE method often had superior statistical size and power. For the case where $\pi_{r0} = \pi_{s0} = 70\%$, the two procedures had identical operating characteristics. The Simon–Schultz method suffered from the same deficiencies as described earlier, leading to designs with poor operating characteristics.

The Opt-CoE procedure was superior to the other procedures (by expected sample sizes and having design parameters closer to the desired levels) in all the settings examined except for the case where $\pi_{r0} = \pi_{s0} = 0.70$; in this case, all the designs had identical operating characteristics.

Table 5. Operating characteristics for three methods: Simon's adjusted design (or a modified Simon–Schultz procedure), a flexible co-primary-binomial method evaluated at Simon's sample sizes (Flx-CoE), and an optimal co-primary-binomial method (Opt-CoE)

Method	π_{r0} (%)	π_{s0} (%)	n_1	n	PET (%)	$E[N_t H_0]$	Size (%)	Power H_r (%)	Power H_s (%)
Simon ^a	5	5	8	23	44	16.4	4.6	83.1	83.1
Flx-CoE					44	16.4	4.6	83.1	83.1
Opt-CoE			7	23	48	15.2	4.4	81.1	81.1
Simon ^a	10	10	10	38	54	22.8	5.0	84.0	84.0
Flx-CoE					26	30.8	3.9	84.3	88.2
Opt-CoE			11	30	49	20.8	4.6	80.6	80.6
Simon ^a	20	20	13	55	56	31.4	5.5	83.7	83.7
Flx-CoE					37	39.3	4.7	87.0	86.9
Opt-CoE			17	45	57	28.9	4.2	81.0	81.0
Simon ^a	30	30	17	65	60	36.2	4.7	83.2	83.2
Flx-CoE					46	42.8	4.1	86.2	85.5
Opt-CoE			18	50	52	33.3	4.3	80.0	80.0
Simon ^a	50	50	18	57	58	34.5	5.0	82.9	82.9
Flx-CoE					45	39.4	4.2	82.3	84.3
Opt-CoE			20	51	56	33.6	4.2	80.2	80.2
Simon ^a	70	70	12	35	55	22.2	4.6	82.9	82.9
Flx-CoE					55	22.2	4.6	82.9	82.9
Opt-CoE			12	35	55	22.2	4.6	82.9	82.9
Simon ^{b,c}	10	15	12	45	21	38.1	3.9	90.4	88.0
Flx-CoE					48	29.0	3.5	86.7	83.9
Opt-CoE			17	35	58	24.6	4.4	84.7	80.8
Simon ^b	40	50	16	61	8	57.6	5.1	89.9	88.2
Flx-CoE					43	41.7	4.5	84.8	85.2
Opt-CoE			20	54	57	34.8	4.2	81.2	80.1
Simon ^b	50	65	18	57	21	48.9	4.5	87.3	92.5
Flx-CoE					49	37.9	4.1	83.8	90.7
Opt-CoE			19	47	57	30.9	4.7	83.0	80.0

PET: probability of early termination.

Note that $\Delta_r = \Delta_s = 0.20$, $\alpha = 0.05$, $\beta = 0.20$.

^aProcedure utilized the modified Simon procedure as defined earlier.

^bProcedure utilized the modified Simon–Schultz method.

^cThe Simon–Schultz method recommended always proceeding to the second stage. In this case, we required at least one response before proceeding to the second stage.

Discussion

General points

The proposed method discussed here has been utilized in a number of Phase II trials by the Gynecologic Oncology Group (GOG). During the development of GOG 0229E, there were discussions about the cytotoxic attributes of bevacizumab and suggestions of replacing 6-month PFS with response. This seemed appropriate since a sufficient number of responses were seen in a Phase II ovarian cancer study to justify further study [21]. However, some investigators were hesitant about its degree of cytotoxic activity (with a 21% observed response rate) and the sensitivity of the trial to detect activity through other mechanisms (e.g., 40% of the patients had 6-month PFS). A nice solution to these problems is to incorporate both outcomes into the design as a co-primary end point study. The probability of a type I error can be controlled without

causing undue costs to statistical power or sample size requirements. The gain was the assurance of trial sensitivity to two mechanisms of drug activity.

An important characteristic of the proposed design (as well as the one proposed by Yu) is its robustness against the degree of association between the co-primary end points and, moreover, the conservative quality of the assumption of independence of the two end points, as indicated at the end of the section titled 'Searching for flexible designs that obtain the planned α , β_r , and β_s ' [13]. This feature is important because the level of association likely changes in significant ways from one study to the next. Therefore, designing a study (with a method that is not robust) based on a particular level of association is unlikely to yield reliable conclusions.

Unfortunately, there is a degree of dependence with regard to the PET. In the illustrations provided here, the PET varied from about 43%–54%. These

values are also less than typically seen with Simon's univariate designs. This leads to more trials that complete the second stage with truly inactive agents. In this regard, the trials are a bit less efficient than single end point trials even if the overall risk of recommending an inactive drug remains the same.

Differences with other works

Lin *et al.* [14] developed a procedure that utilizes the primary end point at the interim analysis and then uses both end points at the end of the trial. This is an important distinction. Our procedures use both end points at both stages of the trial since we were equally interested in either efficacy measure. The methods of Lu *et al.* [10] can potentially be applied to the problem discussed here if all (or at least most) the patients who respond have a 6-month PFS. The applicability depends on the duration of response in most people. The data within the GOG indicated that patients with short-term responses were not exceedingly rare, so this category of patients was preferably modeled. Lin and Chen [11] transformed bivariate information (complete responses and partial responses) into a univariate scale, utilizing a weighted linear combination with greater weight given to complete responses. Since these characteristics are mutually exclusive, and because we were equally interested in activity on either scale, we did not examine this approach in great detail. We believe that there is no reason why the methods of Bryant and Day [6] or Thall and Cheng [7] could not be applied to problems of efficacy by substituting non-toxic reactions with beneficial response. However, their decision criteria would require activity on both end points (or at least noninferiority on one dimension) before declaring the drug active. There are cases where such decision criteria are required (e.g., by the Food and Drug Administration (FDA) in drugs being marketed for activity against migraines), but this was not required for our purposes. For similar reasons, the methods of Yu *et al.* [13] and Conaway and Petroni [8] were not interesting to us. The methods proposed by Thall *et al.* appear more suitable for investigations that can monitor patients sequentially [15]. Stallard *et al.* [16] provide a Bayesian alternative that can include additional criteria for conducting a Phase III study such as its cost and potential benefit to future patients. Many of the methods provided here follow an unpublished technical report by Sill and Yothers [22].

Comparison of methods

First, the PET was exceptionally low for the Simon-Schultz procedure, making the utilization of an interim analysis almost unworthy of incorporation

into the design (7% for the case where $\pi_{r0} = 0.10$ and $\pi_{s0} = 0.25$ in Table 4). This low PET resulted mostly from Schultz's procedure. Because the univariate procedures (especially Simon) expect a higher percentage of the trials to be erroneously stopped early under the alternative hypothesis, they are designed to have more 'generous' thresholds in the second stage. When this normally univariate procedure is tied into a co-primary procedure, the higher than expected probability of proceeding to a second stage results in higher than expected power. These designs are therefore systematically overpowered. This procedure is not recommended for the design of co-primary studies.

Simon's procedure provides a formidable competitor to the flexible procedure (Flx-CoE) in the case where $\pi_{r0} = \pi_{s0}$ and $\Delta_r = \Delta_s$. However, this method requires fairly strong assumptions that may not apply to many clinical questions. In addition, if the targeted sample size is not met in either stage, the method does not offer any remedial recommendations. Finally, if a sponsor can attain a sample size to such a rigorous requirement, then they will generally do better by applying an Opt-CoE procedure.

Future work

The flexibility of the design has been questioned by several statisticians in the field as a clever procedure to enable investigators to test the null hypothesis multiple times in an attempt to obtain significant results. Although abuse of the method in this way is possible as it is with Chen and Ng's method [4], the procedure should only be used by organizations that have relatively imprecise control over study accrual, and the decision rules should only be applied once at each stage. For organizations that have precise control over study accrual, a rigid design should be used. They should use a method such as that provided by the 'optimum' procedure described in section 'Comparisons with other procedures', which will be posted on our website. Also, we are looking into developing a truly exhaustive algorithm that finds the optimal or minimax designs for co-primary rigid designs.

Acknowledgments

The authors would like to thank the referees and editors for their helpful comments that improved this article. The authors thank Bill Brady at Roswell Park Cancer Institute and the GOG for his thoughts, insights, and study on these designs using SAS macros.

Funding

The research by Michael W Sill was supported in part by the National Cancer Institute grant to the Gynecologic Oncology Group Statistical and Data Center (CA37517). The research by Samuel

Litwin was supported in part by NIH grant P30 CA 06927, NCI FCCC-UPenn ovarian cancer SPORE P50 CA083638, and an appropriation from the Commonwealth of Pennsylvania.

Conflict of interest

None

References

1. Gehan EA. The determination of the number of patients required in a preliminary and a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis* 1961; **13**: 346–53.
2. Fleming TR. One-sample multiple testing procedure for phase II clinical trials. *Biometrics* 1982; **38**: 143–51.
3. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials* 1989; **10**: 1–10.
4. Chen TT, Ng TH. Optimal flexible designs in phase II clinical trials. *Stat Med* 1998; **17**: 2301–12.
5. Green SJ, Dahlberg S. Planned versus attained design in phase II clinical trials. *Stat Med* 1992; **11**: 853–62.
6. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics* 1995; **51**: 1372–83.
7. Thall PF, Cheng SC. Optimal two-stage designs for clinical trials based on safety and efficacy. *Stat Med* 2001; **20**: 1023–32.
8. Conaway MR, Petroni GR. Bivariate sequential designs for phase II trials. *Biometrics* 1995; **51**: 656–64.
9. Zee B, Melnychuk D, Dancey J, Eisenhauer E. Multinomial phase II cancer trials incorporating response and early progression. *J Biopharm Stat* 1999; **9**: 351–63.
10. Lu Y, Jin H, Lamborn KR. A design of phase II cancer trials using total and complete response endpoints. *Stat Med* 2005; **24**: 3155–70.
11. Lin SP, Chen TT. Optimal two-stage designs for phase II clinical trials with differentiation of complete and partial responses. *Commun Stat A-Theor* 2000; **29**: 923–40.
12. Panageas KS, Smith A, Gonen M, Chapman PB. An optimal two-stage phase II design utilizing complete and partial response information separately. *Control Clin Trials* 2002; **23**: 367–79.
13. Yu J, Kepner JL, Iyer R. Exact tests using two correlated binomial variables in contemporary cancer clinical trials. *Biom J* 2009; **51**: 899–914.
14. Lin X, Allred R, Andrews G. A two-stage phase II trial design utilizing both primary and secondary endpoints. *Pharm Stat* 2008; **7**: 88–92.
15. Thall PF, Simon RM, Estey EH. Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Stat Med* 1995; **14**: 357–79.
16. Stallard N, Thall PF, Whitehead J. Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics* 1999; **55**: 971–77.
17. Therasse P, Arbuck SG, Eisenhauer EA, et al. New guidelines to evaluate the response to treatment in solid tumors. European Organization for Research and Treatment of Cancer, National Cancer Institute of the United States, National Cancer Institute of Canada. *J Natl Cancer Inst* 2000; **92**: 205–16.
18. Aghajanian C, Sill MW, Darcy KM, et al. Phase II trial of bevacizumab in recurrent or persistent endometrial cancer: A Gynecologic Oncology Group Study. *J Clin Oncol* 2011; **29**: 2259–65.
19. Aghajanian C, Sill M, Darcy K, et al. A phase II evaluation of bevacizumab in the treatment of recurrent or persistent endometrial cancer: A Gynecologic Oncology Group (GOG) study. *J Clin Oncol* 2009; **27**: 284s (Abstract (ASCO#5531)).
20. Schultz JR, Nichol FR, Elfring GL, Weed SD. Multiple-stage procedures for drug screening. *Biometrics* 1973; **29**: 293–300.
21. Burger RA, Sill MW, Monk BJ, Greer BE, Sorosky JI. Phase II trial of bevacizumab in persistent or recurrent epithelial ovarian cancer or primary peritoneal cancer: A Gynecologic Oncology Group Study. *J Clin Oncol* 2007; **25**: 5165–71.
22. Sill MW, Yothers G. *A method for Utilizing Bivariate Efficacy Outcome Measures to Screen Agents for Activity in 2-stage Phase II Clinical Trials* (Technical Report no. 0608). Buffalo, NY: State University of New York at Buffalo, 2006.