A Method for Utilizing Bivariate Efficacy Outcome Measures to Screen Agents for Activity in 2-Stage Phase II Clinical Trials

Michael W. Sill[1] and Greg Yothers[2]

[1]Adjunct Instructor; Research Assistant Professor, SUNY at Buffalo
Senior Biostatistician, GOG Statistical and Data Center
Roswell Park Cancer Institute; Elm and Carlton Streets; Buffalo, NY 14263

[2] Research Assistant Professor, Dept. of Biostatistics, University of Pittsburgh
Coordinating Statistician for Colorectal Trials, NSABP Biostatistical Center
201 N. Craig Street, Suite 350, Pittsburgh, PA 15213

Abstract: A general method is proposed to evaluate drugs with two dichotomous measures of efficacy with the ability to detect activity on either scale 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 scales. The method provides a flexible 2-stage design enabling early closure of the study when initial results indicate the drug is not promising. The design is proposed in the context of clinical trials involving cancer research where tumor response is typically used to evaluate cytotoxic agents and progression-free survival status at 6 months is used to evaluate cytostatic agents. These ideas can be easily broadened to more general applications where investigators are equally interested in detecting drug activity in either of two dimensions. Because the design is flexible with regard to actual accrual numbers, its use by cancer cooperative groups and other nationally based studies should be appealing. Procedures are available in SAS for implementing these methods.

Key Words: Binomial distribution, multinomial distribution, correlated primary endpoints.

Introduction:

Many phase II clinical trials examining anti-cancer drugs use tumor response as the primary endpoint of the study. There are several reasons why this surrogate is attractive. First, because many agents are designed to kill tumor cells (i.e. drugs that are cytotoxic), it is natural to expect a reduction of the tumor burden within the patient over time that is detectable upon radiographic examination. Agents with more cytotoxic activity are expected to produce greater reductions in tumor burden within patients and a higher proportion of patients with objective responses, and this beneficial effect is expected to translate into longer survival. Another advantage to using tumor response is the ability to determine the outcome for most patients relatively quickly. In many cases, results are obtained within 4 months of study entry. This enables relatively quick evaluation of experimental agents in the phase II setting. Finally, some physicians argue that tumor response is not simply a surrogate for survival but indeed leads to a better quality of life by reducing the level of pain experienced by some patients. This measure of anti-tumor activity has become a standard for evaluating agents in diseases with solid tumors.

With the development of cytostatic agents, on the other hand, tumor response became an inadequate measure of drug activity because these drugs were designed to stabilize tumor growth, not actively kill existing tumor cells. If tumor growth could at least be arrested, patients were expected to survive longer from a lack of deteriorating health status. The idea behind these drugs is to make cancer a “chronic” condition if a cure is not available. However, it is possible to conceive of a drug capable of this kind of activity without producing any objective responses, so another surrogate variable was clearly needed to enable early detection. The GOG responded with the use of a dichotomized variable based on the patient's progression-free survival status at 6 months of study entry. If the patient survived progression-free for at least 6 months, then the patient was deemed a successful outcome. Otherwise, the patient was deemed a treatment failure.

Progression-free survival is an attractive surrogate because it has been shown to be associated with survival, and it is capable of detecting stabilization of disease. There are also large amounts of historical data that can be used to establish interesting and uninteresting probabilities of surviving progression-free beyond any point in time. However, there are some disadvantages with using this endpoint. Because the endpoint requires the establishment of the patients' status 6 months after study entry, it often takes longer to determine the outcome in comparison to tumor response. Given the realities of patient care and trial management, it is not unusual for institutions to report this outcome as long as 9 months after the patient enters the study. Lack of familiarity with this endpoint can be problematic as well, but over time, exposure to these kinds of endpoints should lead to greater acceptance.

If the agent under consideration is clearly expected to act solely in a cytostatic manner, then the use of tumor response as a surrogate measure is obviously inappropriate. However, some agents may have in vitro indications of possessing both cytostatic as well as cytotoxic activity, so it is conceivable that arguments can be made to use tumor response because of its familiarity and shorter time of determination. Yet, if the agent is more cytostatic than cytotoxic, an important and useful drug may be overlooked in the screening process.

One solution to this problem is to incorporate both variables as primary endpoints to be evaluated simultaneously. If the agent possesses a fairly high degree of cytotoxic activity, then this agent should be detectable with tumor response in a timely manner. If on the other hand the agent possesses mostly cytostatic activity, then the agent should not escape detection because the study is also design to screen for these types of agents as well. This provides the motivation for the proposed design.

It is important to remember that both measures are not mutually exclusive, nor are they assumed to be independent. Because the level of association between the outcome measures often depend on the type of treatment administered, it is not feasible to make any assumptions regarding the level of association when designing the study. Fortunately, the operating characteristics of these designs are not heavily dependent on the level of association. One of the keys to creating a successful study design is knowing that the probability of a type I error is highest when the outcome measures are independent and that the probability of a type II error is highest when the outcome measures have a high degree of association.

Methodology:

The null hypothesis can be formulated as follows: H_0: $\pi_r \leq \pi_{r0}$ and $\pi_s \leq \pi_{s0}$, where π_r is the true response rate and π_s is the true proportion of patients who survive progression-free for at least 6 months. π_{r0} and π_{s0} are specified values obtained from historical data that are believed to be uninteresting. The alternative hypothesis is the complement of this null parameter space.

Test Statistics and Distributions
Before a testing procedure can be constructed and characterized, consideration must be given to the joint distribution of the number of patients who respond or survive progression-free for 6 months. The parameters of this distribution can be conveniently displayed in the following table:

Table 1

		PFS>6Mo.		
		Yes	No	
Response	Yes	π_{11}	π_{12}	π_r
	No	π_{21}	π_{22}	π_{2+}
		π_s	π_{+2}	

Similarly, the number of patients who have the qualities of interest in this particular trial can be tabulated in the following table:

Table 2

		PFS>6Mo.		
		Yes	No	
Response	Yes	$X_{11(k)}$	$X_{12(k)}$	$X_{r(k)}$
	No	$X_{21(k)}$	$X_{22(k)}$	$X_{2+(k)}$
		$X_{s(k)}$	$X_{+2(k)}$	n(k)

where n(k) is the total sample size for stage k=1,2 of the trial, $X_{r(k)}$ is the total number of patients who have an objective response in stage k, and $X_{s(k)}$ is the total number of patients in the same stage who survive progression-free for 6 months. It can be shown that the distribution of $X_{ij(k)}$ where i=1,2 and j=1,2 is a 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 of this distribution can be written as follows:

$$f\left(x_{ij(k)}\right) = \frac{n(k)!}{x_{11(k)}!\,x_{12(k)}!\,x_{21(k)}!\,x_{22(k)}!}\pi_{11}^{x_{11(k)}}\pi_{12}^{x_{12(k)}}\pi_{21}^{x_{21(k)}}\pi_{22}^{x_{22(k)}} \qquad (1)$$

The null hypothesis will be accepted 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 critical values for stage 1, $X_r = X_{r(1)} + X_{r(2)}$, $X_s = X_{s(1)} + X_{s(2)}$, and C_r and C_s are the critical values for the cumulative number of patients who have a response or survive progression-free for 6 months at the end of stage 2. Note that the following relationships hold in general:

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

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

$$X_{22(k)} = n(k) - X_{11(k)} - X_{12(k)} - X_{21(k)}$$

To determine the probability of accepting the null hypothesis after a particular stage using (1), it is helpful to define a probability mass function and a cumulative distribution function in terms of n(k), $X_{s(k)}$, and $X_{r(k)}$:

$$P\left(X_{s(k)} = x_{s(k)}, X_{r(k)} = x_{r(k)}\right) = f\left(n(k), x_{s(k)}, x_{r(k)}\right) = \sum_{X_{11(k)}=\max\left\{0, x_{r(k)}+x_{s(k)}-n(k)\right\}}^{\min\{x_{r(k)}, x_{s(k)}\}} f\left(x_{ij(k)}\right)$$

$$P\left(X_{s(k)} \le s, X_{r(k)} \le r\right) = F\left(n(k), s, r\right) = \sum_{X_{r(k)}=0}^{r} \sum_{X_{s(k)}=0}^{s} \sum_{X_{11(k)}=\max\left\{0, X_{r(k)}+X_{s(k)}-n(k)\right\}}^{\min\{X_{r(k)}, X_{s(k)}\}} f\left(x_{ij(k)}\right)$$

The probability of early termination (PET), which is the probability of accepting the null hypothesis after the first stage can be calculated simply with the cumulative distribution function and using the first stage parameters, i.e.,

$$PET = F(n(1), C_{s(1)}, C_{r(1)})$$

where n(1) is the first stage sample size. In order for the null hypothesis to be accepted after the second stage, it is required that the outcome after the first stage not lie within the acceptance region (i.e. $X_{s(1)} \le C_{s(1)}$ and $X_{r(1)} \le C_{r(1)}$) but the outcome in the second stage lie within its acceptance region (i.e. $X_s \le C_s$ and $X_r \le C_r$). In order for this condition to be true, it is required that the following condition hold: $X_{s(1)} > C_{s(1)}$ or $X_{r(1)} > C_{r(1)}$, and simultaneously that $X_{s(1)} \le C_s$ and $X_{r(1)} \le C_r$ for the first stage outcome. Using Figure 1, this region corresponds to the union of Region B and Region C (Note that it is possible for C_s and C_r to be greater than n(1)).

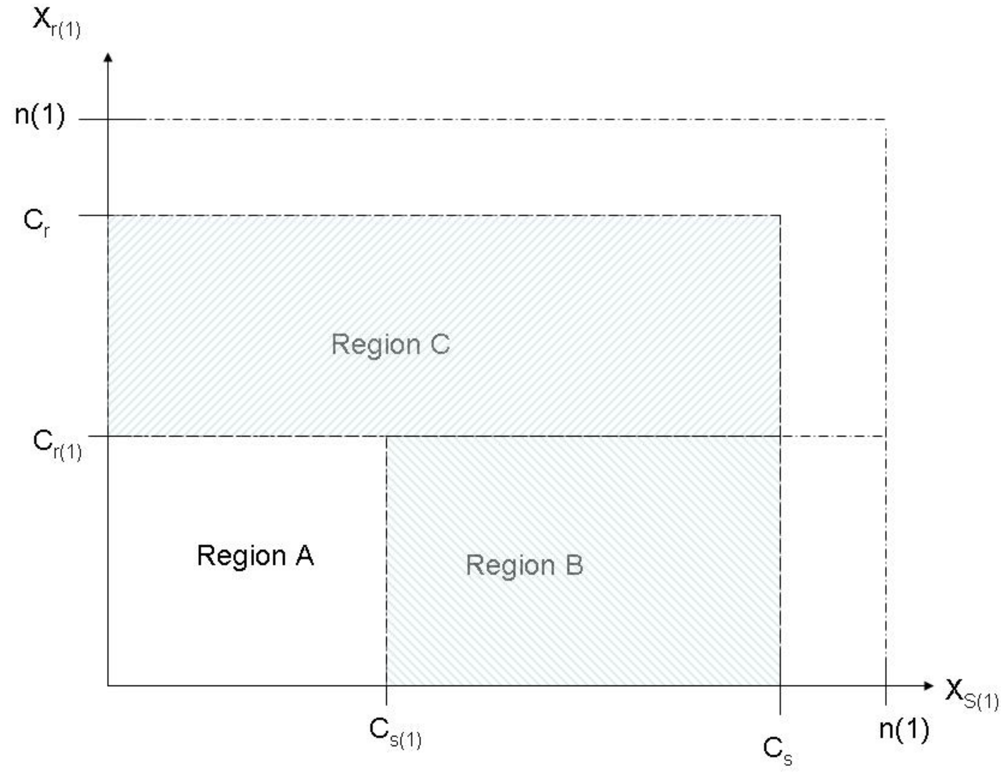

Figure 11.1: The sample space for the number of patients who have responses and survive progression-free for 6 months in stage 1.

To calculate the probability that the null hypothesis is accepted in the second stage, it is important to note that $X_s – X_{s(1)}$ and $X_r – X_{r(1)}$ are marginal totals equal to $X_{s(2)}$ and $X_{r(2)}$ 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 $X_r – X_{r(1)} \leq C_r – X_{r(1)}$. Using this fact along with the distributional fact provided in the previous sentence as well as the restrictions derived above, the probability of accepting the null hypothesis in stage 2 is calculated as:

$$\sum_{X_{s(1)}\in B\cup C}\ \sum_{X_{r(1)}\in B\cup C} f\left(n(1), X_{s(1)}, X_{r(1)}\right) F\left(n-n(1), C_S - X_{s(1)}, C_r - X_{r(1)}\right)$$

where $B \cup C$ is Region B union Region C and $n = n(1) + n(2)$. The total probability of accepting the null hypothesis is the sum of the probabilities of accepting the null hypothesis in each stage.

A good design has a high probability of early termination and accepting the null hypothesis overall when it is indeed the case. We are also interested in detecting agents that are active either cytostatically or cytotoxically. So, a good design is also one that has a low probability of accepting the null hypothesis when $\pi_r = \pi_{r0} + \Delta_r$ or $\pi_s = \pi_{s0} + \Delta_s$

where π_{r0} is the largest response rate in the null parameter space, Δ_r is a clinically significant increase in the response rate, π_{s0} is the largest proportion of patients surviving progression-free for 6 months in the null parameter space, and Δ_s is a clinically significant increase in this proportion. Therefore, there are three characteristics of the design that are of interest to us:

$$\alpha = P\left(reject\, H_0 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0}\right)$$
$$\beta_r = P\left(accept\, H_0 \mid \pi_r = \pi_{r0} + \Delta_r, \pi_s = \pi_{s0}\right)$$
$$\beta_s = P\left(accept\, H_0 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0} + \Delta_s\right)$$

Generally speaking, all of these quantities are ideally kept as small as possible. For a particular set of critical boundaries, $C_{s(1)}$, $C_{r(1)}$, C_s, and C_r, the following quantities can be defined:

$$PETHO = P\left(accept\, H_0\, Stage 1 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0}\right)$$
$$PETHR = P\left(accept\, H_0\, Stage 1 \mid \pi_r = \pi_{r0} + \Delta_r, \pi_s = \pi_{s0}\right)$$
$$PETHS = P\left(accept\, H_0\, Stage 1 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0} + \Delta_s\right)$$
$$TPRTHO = P\left(accept\, H_0 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0}\right)$$
$$TPRTHR = P\left(accept\, H_0 \mid \pi_r = \pi_{r0} + \Delta_r, \pi_s = \pi_{s0}\right)$$
$$TPRTHS = P\left(accept\, H_0 \mid \pi_r = \pi_{r0}, \pi_s = \pi_{s0} + \Delta_s\right)$$

where "PET" stands for probability of early termination and "TPRT" stands for total probability of rejecting the treatment. Given a particular first stage sample size, the first stage critical boundaries, $C_{s(1)}$ and $C_{r(1)}$, are determined by maximizing PETHO among all cases for which PETHR $\leq \beta_r / 2$ and PETHS $\leq \beta_s / 2$ under the assumption of independence for responding and surviving progression-free for 6 months. If the set of critical boundaries are not unique, then the set with the smallest value of $C_{r(1)}$ is selected as the critical boundary. Once the first stage critical boundaries are determined, the second stage critical boundaries are determined so that the following cost function is minimized for a particular sample size (under the assumption of independence):

$$C = (1 - TPRTHO)^2 + (TPRTHR)^2 + (TPRTHS)^2$$

Again, if the set of critical boundaries are not unique, then the set with the smallest value of C_r is selected as the critical boundary.

Decision Rule: If either $X_{s(1)} > C_{s(1)}$ or $X_{r(1)} > C_{r(1)}$ after the first stage, then the study will open to a second stage of accrual to further evaluate the activity of the drug. If either $X_s > C_s$ or $X_r > C_r$ after the second stage and clinical judgment indicates, then the agent will be deemed clinically interesting and worthy of further investigation.

<u>Illustration</u>:

The design parameters for GOG 0170I arose from considering the results of a series of protocols in GOG 0126 and GOG 0146. The patients entered on GOG 0126 are required to have platinum resistant disease (i.e., a platinum free interval of less than 6 months) while the patients entered on GOG 0146 are required to be platinum sensitive (i.e., a platinum-free interval between 6 and 12 months). The GOG allows patients from both populations to enter GOG 0170. Of the 70 patients enrolled onto this protocol in sections C, D, and E for which previous history is currently available, 28 (40%) were platinum sensitive and 42 (60%) were platinum resistant as defined above.

The GOG protocols listed in Table 3 were selected for historical controls because the population is similar to GOG-0170 and the agents investigated were deemed to have minimal activity. The purpose of GOG 0170I is to screen the agent for any cytostatic or cytotoxic activity above this minimum threshold. The number of evaluable patients was adjusted to reflect a change in its definition, which now includes all patients who take any study agent.

Table 3: Measures of Efficacy by Protocol and Section

Protocol	N of evaluable patients	Prob (PFS > 6 months)*	Number of responses (%)
126-B	26	0.15 (0.07)	3 (12%)
126-C	33	0.27 (0.08)	3 (9%)
126-D	27	0.19 (0.07)	2 (7%)
126-E	58	0.17 (0.05)	5 (9%)
126-G	27	0.07 (0.05)	1 (4%)
126-H	26	0.04 (0.04)	1 (4%)
126-K	23	0.22 (0.09)	1 (4%)
146-B	30	0.21 (0.08)	3 (10%)
146-E	23	0.13 (0.07)	1 (4%)
146-F	29	0.29 (0.09)	2 (7%)
146-J	30	0.07 (0.05)	0 (0%)

* Standard error of the estimate for surviving progression-free for 6 months is provided in the parentheses.

Based on Table 3, we can formulate the null hypothesis as follows, H_0: $\pi_r \leq 0.10$ and $\pi_s \leq 0.15$. With $\Delta_r = 0.15$ and $\Delta_s = 0.20$ considered clinically significant, the targeted accrual for the first stage was set to 23 eligible and evaluable patients. The cumulative targeted accrual for the second stage was set to 52 eligible and evaluable patients. Critical values for each stage are provided in the tables below:

Table 4: Critical values for number of responses and PFS at 6 months after Stage 1.

Stage 1					
n(1)	21	22	23	24	25
$C_{r(1)}$	2	2	2	2	2
$C_{s(1)}$	3	4	4	4	5

Table 5: Critical values for number of responses and PFS at 6 months after Stage 2.

Stage 2 (C_s, C_r)					
n(1)\ n	50	51	52	53	54
21	(12,8)	(12,8)	(12,8)	(13,8)	(13,8)
22	(12,8)	(12,8)	(12,8)	(12,8)	(13,8)
23	(12,8)	(12,8)	(12,8)	(12,8)	(13,8)
24	(12,8)	(12,8)	(12,8)	(13,8)	(13,8)
25	(12,8)	(12,8)	(12,8)	(12,8)	(13,8)

The operating characteristics of the decision rules are provided below using the usual definition of power as:

$$Power = 1 - P\left(\text{Accept H}_0\right)$$

Table 6: Average Power of Testing Procedure assuming a uniform distribution over all accrual combinations when accrual guidelines are met in Tables 4 and 5 and under the assumption that response and surviving progression-free for 6 months are independent events:

		PFS > 6 Mo.	
		$\pi_S = 0.35$	$\pi_S = 0.15$
Response	$\pi_r = 0.25$	99%	91%
	$\pi_r = 0.10$	93%	9%

Table 7: Average Power of Testing Procedure assuming a uniform distribution over all accrual combinations when accrual guidelines are met in Tables 4 and 5 and under the assumption that response and surviving progression-free for 6 months are not independent events:

		PFS > 6 Mo.	
		$\pi_S = 0.35$	$\pi_S = 0.15$
Response	$\pi_r = 0.25$	96%	90%
	$\pi_r = 0.10$	91%	8%

To assess the operating characteristics when the two primary endpoints are not independent, the probability calculations (carried out as outlined above) were done with the assumption that the joint probability was $\pi_{11} = 0.90 \min\{ \pi_r, \pi_s \}$, which would carry a fairly high degree of association.

The average probability of early termination when the null hypothesis is true is 43% under independence and 53% under high association.