Statistics and Its Interface Volume 5 (2012) 463–469

Efficient designs for phase II oncology trials with ordinal outcome

Anastasia Ivanova*, Jane Monaco and Thomas Stinchcombe

Phase II trials in oncology are usually single arm studies to screen oncology treatments based on tumor response. Treatment response can be categorized into one of four types: complete response, partial response, stable disease and progressive disease. Investigators usually dichotomize responses in phase II trials and use a simple hypothesis regarding that binary outcome. We describe an efficient design to test an intersection hypothesis where the drug is not considered promising if both tumor response (defined as complete or partial response) and disease control (defined as tumor response or stable disease) are low. The new design can be generated using easy-to-use software that is available at http://cancer.unc.edu/biostatistics/program/ivanova/.

Keywords and phrases: Simon's design, Phase II trial, Ordinal outcome, Tumor response.

1. INTRODUCTION

In phase II oncology trials, investigators are interested in screening potential treatments for efficacy often based on four mutually exclusive categories defined by Response Evaluation Criteria in Solid Tumors (RECIST) [3]: Complete Response (CR), Partial Response (PR), Stable Disease (SD) or Progressive Disease (PD). Simon [20] proposed a two-stage design for single-arm trials with binary outcome which has become widely employed. Using this method, a trial may be stopped after the first stage if the agent shows insufficient promise. This two-stage design offers ethical and financial benefits compared to single-stage designs because ineffective agents are abandoned earlier and therefore in general fewer patients are required. Simon's design is often applied to tumor response (TR) defined as either CR or PR. More recently, disease control (DC) defined as CR, PR or SD, has been used as a primary outcome [1, 2].

Our motivating example is a trial in patients with advanced non-small cell lung cancer who have experienced disease progression after platinum-based therapy. The tumor response (TR) rate to currently available second-line therapies is 5–10% [18, 19, 9]. The improvement in overall survival and quality of life observed with these therapies is most likely related to disease control given the modest tumor response rates. In situations such as this, when TR is modest, clinicians may be interested in disease control (DC) since it may better predict clinical outcome. Interest has increased in novel endpoints for evaluating new therapeutic agents in oncology [6]. Historically the TR was a frequently used endpoint to assess novel cytotoxic agents which induced tumor regression, but many targeted agents are cytostatic and inhibit tumor growth rather than cause tumor regression. Issues in determining tumor response have included variation in the assessment of response [4] or variability in tumor measurements on repeat imaging [15]. An unblinded single arm trial may be susceptible to bias in the assessment of response by the investigator and lacks a "control arm" for comparison. However, tumor response remains an important clinical event in order to identify biomarkers for future development [13, 16]. Thus, both the tumor response (TR) and disease control (DC) have clinical value and aid in the development of novel agents. Therefore testing both tumor response (TR) and disease control (DC) formally in a phase II trial has important clinical implications.

Several authors have addressed the issue of testing potential drugs in phase II trials based on different outcomes and different configurations of the hypotheses. Panageas [17] proposed a method that rejects the treatment if it does not achieve desirable CR or PR rate. Specifically, they tested no complete response effect (H_{0C}) and no partial response effect (H_{0P}) individually using a trinomial distribution, but did not test the tumor response defined as CR + PR. Lin and Chen [11] developed an extension that weighted the two outcomes relative to their importance. Testing two binary endpoints simultaneously was considered by Lu et al. [12] and Lin et al. [10] with the latter using Monte Carlo procedures to obtain decision boundaries. Lu et al. [12] considered CR and TR, while Lin et al. [10] considered TR and DC as endpoints of interest. We will consider tumor response (TR) and disease control (DC) as these are frequently of interest, including in our motivating example. However, the proposed method can be used for hypothesis testing based on any ternary outcome, for example, {CR, PR, SD + PD} in place of {CR + PR, SD, PD} (Table 1).

Let p_T and p_D, $p_T \leq p_D$, denote the probability for tumor response and disease control in the population, respectively, and p_{0T} and p_{0D} denote the null probabilities of tumor response and disease control. Consider the following simple

*Corresponding author.

Table 1. Response outcomes in phase II cancer trials

Complete Response, CR	Tumor Response, TR = CR or PR	Disease Control, DC = TR or SD = CR or PR or SD
Partial Response, PR		
Stable Disease, SD	No Response = SD or PD	
Progressive Disease, PD		Progressive Disease, PD

hypotheses:

$$H_{0T} : p_T \leq p_{0T} \text{ versus } H_{1T} : p_T > p_{0T},$$
$$H_{0D} : p_D \leq p_{0D} \text{ versus } H_{1D} : p_D > p_{0D}.$$

Lu et al. [12] considered testing an intersection hypothesis $H_{0T} \cap H_{0D}$. While Lu et al. [12] considered CR and TR rather than TR and DC, we use TR and DC when referring to their method without loss of generality. We will refer to their method as the LJL design. According to the LJL design, we can test $H_{0T} \cap H_{0D}$ with the objective to accept the treatment based on either promising TR or promising DC rates. The treatment is not considered promising if both TR and DC rates are low. Let α be the type I error rate for testing $H_{0T} \cap H_{0D}$. Let p_{AT} and p_{AD} be the values for the alternative hypothesis for TR and DC rates respectively, and β, β_T, β_D be the type II error rates associated with testing $H_{0T} \cap H_{0D}$ given $\{p_T = p_{AT}$ and $p_D = p_{AD}\}$, $p_T = p_{AT}$ and $p_D = p_{AD}$ respectively. The LJL method considered designs such that 1) treatment is concluded to be effective with probability of at most α when $p_T = p_{0T}$ and $p_D = p_{0D}$, 2) for $p_T = p_{AT}$ and any p_D the probability to accept the treatment is at least $1-\beta_T$, 3) for any p_T and $p_D = p_{AD}$ the probability to accept the treatment is at least $1 - \beta_D$. Note that H_{0T} and H_{0D} were not tested but rather it was required to have good power for testing $H_{0T} \cap H_{0D}$ when $p_T = p_{AT}$ and good power for testing $H_{0T} \cap H_{0D}$ when $p_D = p_{AD}$. Since the focus was on achieving given power when $p_T = p_{AT}$ or $p_D = p_{AD}$, the rejection regions considered in the LJL method and the method in [10] were restricted to the regions with linear boundaries t and d, such that $\{(X_T, X_D) : X_T > t \text{ or } X_D > d\}$, where X_T and X_D are the number of patients with tumor response and with disease control in the trial and t and d are constants that define rejection region.

There exist many two-stage designs for given power and type I error rates. Simon [20] tabulated the designs that minimize either the maximum sample size ("minimax" design) or the expected sample size under the null hypothesis ("optimal" design). Jung et al. [8] proposed minimizing the weighted average of the maximum sample size and the expected sample size under the null. These admissible designs feature the minimax and optimal designs as special cases and often yield other good designs. Because the optimal and minimax designs can often result in first stage sample sizes which are too small or too large, having other options provided by admissible designs is desirable. In our experience, an investigator often chooses an admissible design with preferable stage one sample size over optimal or minimax designs. Our proposed method computes all admissible designs.

We extend the method in [12] in several ways. First, we propose a two-stage design that tests $H_{0T} \cap H_{0D}$ as the main objective without restrictions on the shape of rejection region. Second, our approach allows specifying desirable β, the type II error rate, to test $H_{0T} \cap H_{0D}$ given promising rates of TR and SD, as well as β_T and β_D. Third, the proposed way of ordering points in the sample space allows developing an efficient algorithm to search over possible designs for each total sample size n and stage 1 sample size n_1. This algorithm is implemented in our easy-to-use web-based software to obtain all admissible designs. Fourth, we describe how to test H_{0T} and H_{0D} individually as well as testing $H_{0T} \cap H_{0D}$. Our method yields higher power, lower expected sample size under the null and lower maximum total sample size, to test $H_{0T} \cap H_{0D}$ compared to the method of Lu et al. [12]. The power when $p_T = p_{AT}$ or when $p_D = p_{AD}$ is higher than the power in [12] as well in many cases.

The paper is organized in the following way. Section 2 describes the formulation of the hypotheses and describes the rejection region for a single stage design, including an example. The two-stage design with an example is described in Section 3. The proposed method is extended in Section 4 to include individual tests for H_{0T} and H_{0D}. The application to the phase II trial in non-small cell lung cancer and discussion are found in Sections 5 and 6.

2. SINGLE STAGE DESIGN

Consider testing

$$H_0 : H_{0T} \cap H_{0D} \text{ versus } H_1 : H_{1T} \cup H_{1D}.$$

Our goal is for given type I and type II error rates α, β, β_T, and β_D to find a single stage design by determining the total sample size n and futility region S_1, such that

$$\begin{aligned}
&\Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_{0T}, p_{0D}\right\} \leq \alpha, \\
&\Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_{AT}, p_{AD}\right\} \geq 1 - \beta, \\
&\min_{p_D} \Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_{AT}, p_D\right\} \geq 1 - \beta_T, \\
&\min_{p_T} \Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_T, p_{AD}\right\} \geq 1 - \beta_D.
\end{aligned} \tag{1}$$

Applying proposition A.1 from [12], (1) is equivalent to

$$
\begin{aligned}
&\Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_{0T}, p_{0D}\right\} \leq \alpha,\\
&\Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_{AT}, p_{AD}\right\} \geq 1-\beta,\\
&\Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_{AT}, p_D = p_{AT}\right\} \geq 1-\beta_T,\\
&\Pr\left\{(X_T, X_D) \in \bar{S}_1 \mid p_T = 0, p_{AD}\right\} \geq 1-\beta_D.
\end{aligned}
\tag{2}
$$

We now describe how to construct the rejection region $\bar{S}_1$ for testing $H_{0T} \cap H_{0D}$ in a single stage trial. Let $\Pr(X_T, X_D \mid p_T, p_D, n)$ be the probability of outcome (X_T, X_D) after n patients have been assigned and given p_T and p_D. For any given outcome (x_T, x_D), we compute the probability $V(x_T, x_D) = \Pr(X_T \geq x_T \text{ or } X_D \geq x_D \mid p_{0T}, p_{0D}, n)$. We use the values $V(x_T, x_D)$ to order outcomes in the sample space. To form the rejection region, we first take outcomes with the smallest $V(x_T, x_D)$, then add outcomes with the second smallest possible value of $V(x_T, x_D)$, etc. We continue the process until the sum of the probabilities of the outcomes in rejection region given $H_0, \Pr(x_T, x_D \mid p_{0T}, p_{0D}, n)$, is less than α.

For each observed outcome (x_T, x_D), define the p-value as the sum of $\Pr(X_T, X_D \mid p_{0T}, p_{0D}, n)$ for all (X_T, X_D) such that $V(X_T, X_D) \leq V(x_T, x_D)$:

$$
\begin{aligned}
&pv(x_T, x_D)\\
&\quad = \sum_{(X_T, X_D): V(X_T, X_D) \leq V(x_T, x_D)} \Pr(X_T, X_D \mid p_{0T}, p_{0D}, n).
\end{aligned}
\tag{3}
$$

For a single-stage design, the rejection region can be defined as the set of points with p-values less than α. As there are many designs for given α and β, we are usually interested in a single-stage design with the smallest n.

We use an example with a small sample size to illustrate the differences between the LJL and proposed approaches. Clearly, phase II oncology trials have larger total sample sizes.

Example, single stage Figure 1 presents the rejection region for the intersection hypothesis, $H_{0T} \cap H_{0D}$, when $p_{0T} = 0.15$, $p_{0D} = 0.35$, $p_{AT} = 0.55$, and $p_{AD} = 0.75$ with $\alpha = 0.05$. The proposed test rejects $H_{0T} \cap H_{0D}$ for all outcomes shown by dark dots. In comparison, the LJL test for the same α level rejects $H_{0T} \cap H_{0D}$ if $X_T \geq 4$ or $X_D \geq 6$. Points that are in the rejection region of the proposed test and not in the LJL test are marked with crosses. Since power given p_{AT} is computed as the minimum over all possible values of p_D, and the minimum is attained at $p_D = p_{AT}$, both our proposed method and LJL test yield the same power of 0.44 when $p_T = p_{AT}$. Similarly, power for given p_{AD} is computed as the minimum over all possible values of p_T, and the minimum is attained at $p_T = 0$. Both methods yield the same power for DC of 0.61. The power for $H_{0T} \cap H_{0D}$ is much higher for the proposed test: 0.80 versus 0.68 for the LJL test. If TR is the only outcome considered in the trial, then H_{0T} is rejected if $X_T \geq 4$. If DC is the only outcome, the H_{0D} is rejected if $X_D \geq 6$. In this example, the rejection region for the proposed test is actually larger than the union of the rejection regions for H_{0T} and H_{0D}. This is because the proposed test exhausts the α level well compared to tests for a single binary outcome. The actual level attained is 0.047 for the proposed test, 0.040 for the LJL test, 0.01 to test H_T and 0.01 to test H_D.

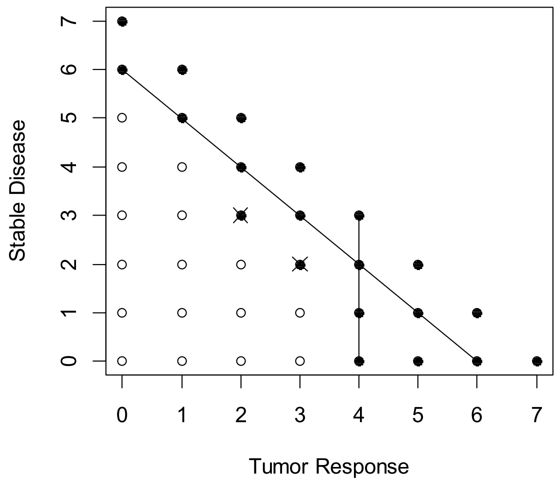

Figure 1. Rejection region for the intersection hypothesis for $p_{0T} = 0.15$, $p_{0D} = 0.35$, $p_{AT} = 0.55$, and $p_{AD} = 0.75$ with $n = 7$ and $\alpha = 0.05$. The proposed test rejects $H_{0C} \cap H_{0T}$ for all outcomes shown by dark dots. Lines show the boundaries for rejection region of the LJL method. The points marked with "×" are additional points in the rejection region using the proposed method.

3. TWO-STAGE DESIGN

3.1 Two-stage design

Let n_1 be the number of patients assigned in stage 1 and $(X_T^{(1)}, X_D^{(1)})$ be the outcome in stage 1. Let the futility region, S_1, be the set of outcomes that do not warrant continuation to stage 2, and let PET be the probability of early termination of the trial under $H_{0T} \cap H_{0D}$,

$$\text{PET} = \Pr\left\{(X_T^{(1)}, X_D^{(1)}) \in S_1 \mid p_{0T}, p_{0D}, n_1\right\}.$$

Region $\bar{S}_1$ that warrants continuation of the trial to stage 2 can be described as the set of outcomes with p-values less than 1 − PET. The process of computing p-value was described in Section 2. Also denote (X_T, X_D) to be the outcome after stage 2, and S_2 to be a set of outcomes for which futility is declared after the trial (fail to reject the null hypothesis $H_{0T} \cap H_{0D}$).

Our goal is for given α, β, β_T, and β_D, to find n, n_1, S_1 and S_2, such that

$$\begin{aligned}
&\Pr\big\{(X_T^{(1)}, X_D^{(1)}) \in \bar{S}_1 \text{ and} \\
&\quad (X_T, X_D) \in \bar{S}_2 \mid p_{0T}, p_{0D}\big\} \leq \alpha, \\
&\Pr\big\{(X_T^{(1)}, X_D^{(1)}) \in \bar{S}_1 \text{ and} \\
&\quad (X_T, X_D) \in \bar{S}_2 \mid p_{AT}, p_{AD}\big\} \geq 1-\beta, \\
&\Pr\big\{(X_T^{(1)}, X_D^{(1)}) \in \bar{S}_1 \text{ and} \\
&\quad (X_T, X_D) \in \bar{S}_2 \mid p_{AT}, p_D = p_{AT}\big\} \geq 1-\beta_T, \\
&\Pr\big\{(X_T^{(1)}, X_D^{(1)}) \in \bar{S}_1 \text{ and} \\
&\quad (X_T, X_D) \in \bar{S}_2 \mid p_T = 0, p_{AD}\big\} \geq 1-\beta_D.
\end{aligned} \tag{4}$$

Similarly to Section 2, we used Proposition A.1 from [12] in (4). Note that a two-stage design can be alternatively described by n, n_1, PET and S_2.

The expected sample size under H_0, EN_0, is defined as

$$EN_0 = n_1 + (n - n_1)\text{PET}.$$

As there might be many designs satisfying criteria (4), we will consider all admissible designs [8], designs that minimize $wEN_0 + (1-w)n$ for some w, such that $0 \leq w \leq 1$. Admissible designs are obtained by an exhaustive search through n, n_1, PET and S_2.

We describe futility region S_i using constants t_i, d_i and a set of points A_i, $i = 1, 2$:

$$\begin{aligned}
S_1(t_1, d_1, A_1) &= \big\{(X_T^{(1)}, X_D^{(1)}) : X_T^{(1)} \leq t_1 \text{ and} \\
&\qquad X_D^{(1)} \leq d_1\big\} \cup A_1, \\
S_2(t_2, d_2, A_2) &= \big\{(X_T, X_D) : X_T \leq t_2 \text{ and} \\
&\qquad X_D \leq d_2\big\} \cup A_2.
\end{aligned}$$

Example, two-stage Figure 2 presents the minimax two-stage design for $p_{0T} = 0.15$, $p_{0D} = 0.35$, $p_{AT} = 0.55$, and $p_{AD} = 0.75$, that attains a power of at least 0.8 ($\beta = 0.2$) for testing $H_{0T} \cap H_{0D}$ with $\alpha = 0.05$. Stage 1 enrolls $n_1 = 5$ patients and the maximum total sample size is $n = 7$. The futility region S_1, the number of responses in $n_1 = 5$ stage 1 patients, is shown by snowflakes. Compared to the LJL method, we do not restrict the shape of the futility region, though in this case the futility region has two linear boundaries. The rejection region after stage 2, $\bar{S}_2$, is shown by dark dots. Futility region S_1can be described by $t_1 = 1, d_1 = 2, A_1 = \{\varnothing\}$ and S_2 by $t_2 = 2, d_2 = 4, A_2 = \{(3,3), (3,4)(0,5), (1,5)\}$. For $n = 7$, there are seven possible designs that attain 0.8 power with type I error rate of 0.05. For comparison, a two-stage design described in [12] does not exist for $n = 7$, and only exists for larger sample sizes.

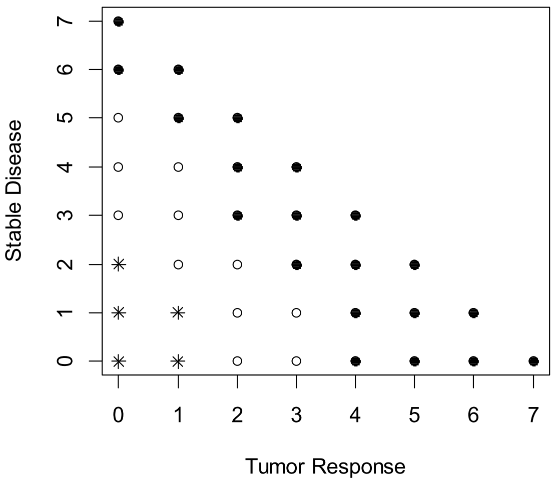

Figure 2. The minimax two-stage design for $p_{0T} = 0.15$, $p_{0D} = 0.35$, $p_{AT} = 0.55$, and $p_{AD} = 0.75$ with $\beta = 0.2$ and $\alpha = 0.05$. Stage 1 enrolls $n_1 = 5$ patients. Snowflakes show the futility region in stage 1. Dark points show rejection region after both stages.

3.2 Results

The new design can be generated using easy-to-use software that we have developed available at http://cancer.unc.edu/biostatistics/program/ivanova/. We illustrate the features of the new approach using examples previously studied ([12], Tables II and IV). Results for two of 24 scenarios in [12] are presented in Table 2. We display all admissible designs including minimax and optimal designs. An admissible design [8] is a design that minimizes $wn + (1-w)EN_0$ for some w in $[0,1]$. A given admissible design in Table 2 minimized the weighted sum over all values of w we report. Wide interval for w indicates that both n and EN_0 are small to yield the smallest $wn + (1-w)EN_0$ for a wide range of w.

We selected a scenario where our approach has the most and the least advantage over the LJL method. We generated the new design when 1) as in [12], only β_T and β_D are given; 2) β_T, β_D and β are given; 3) only β is given. Lu et al. [12] did not use β to construct designs, therefore we used β reported in ([12], Tables II and IV). Since the new design is constructed to have good power for testing $H_{0T} \cap H_{0D}$, the smallest required sample size for the new design when only β is specified is always the same or less. In fact, the proposed method yields a sample size that is strictly less in 21 of 24 scenarios and is equal in the remaining 3. Also across all scenarios and power requirements, the new approach yields smaller EN_0 and much optimal designs with much smaller maximum sample size.

In the first scenario, the sample size for testing $H_{0T} \cap H_{0D}$ is 47, compared to 53 required by the LJL test. When only β_T and β_D are specified, the total sample size reduces from 53 for the LJL test to 42 for the proposed method. We also

Table 2. Comparison of the new approach and the LJL method

$1-\beta$	$1-\beta_T$	$1-\beta_D$	Method	n	n_1	EN_0	w	Design
$p_{0T}=0.01$, $p_{0D}=0.2$, $p_{AT}=0.1$, $p_{AD}=0.4$, $\alpha=0.1$								
–	0.9	0.9	LJL	53	33	39.9		Minimax
			LJL	59	27	37.8		Optimal
–	0.9	0.9	New	42	24	25.8	[0.08, 1]	Minimax
			New	51	20	25.0	[0, 0.08]	Optimal
0.98	0.9	0.9	New	52	30	38.2	[0.77, 1]	Minimax
			New	53	28	35.0	[0.16, 0.77]	
			New	54	27	34.8	[0, 0.16]	Optimal
0.98	–	–	New	47	23	34.0	[0, 1]	Minimax, Optimal
$p_{0T}=0.05$, $p_{0D}=0.2$, $p_{AT}=0.2$, $p_{AD}=0.45$, $\alpha=0.05$								
–	0.6	0.8	LJL	24	16	17.9		Minimax
			LJL	29	12	16.4		Optimal
–	0.6	0.8	New	26	17	19.4	[0.79, 1]	Minimax
			New	27	10	15.8	[0.54, 0.79]	
			New	28	10	14.6	[0.28, 0.54]	
			New	29	11	14.2	[0. 0.28]	Optimal
0.87	0.6	0.8	New	26	17	19.4	[0.78, 1]	Minimax
			New	27	11	15.9	[0, 0.78]	Optimal
0.87	–	–	New	24	14	17.0	[0.41, 1]	Minimax
			New	26	11	15.6	[0, 0.41]	Optimal

observe dramatic reduction in the expected sample size under the null hypothesis, EN_0, by more than 10 patients for each of the inputs 1)–3). In the second scenario, the worst of the 24 scenarios, sample sizes of the minimax designs to test $H_{0T}\cap H_{0D}$ are equal for the two methods. When only β_T and β_D are specified, the sample size for the minimax design for the proposed approach is higher: 26 versus 24 for the LJL design. This is a consequence of the different approaches of the two methods for selection of the points in the rejection region. For example, consider the process of constructing the rejection region to test $H_{0T}\cap H_{0D}$. Let X_S denote the number of patients with stable disease. Using Figure 1, suppose all points $\{(X_T,X_S): X_T\geq 4\}$, except for a point $(4,0)$, and all points $\{(X_T,X_S): X_T+X_S\geq 6\}$ are already in the rejection region. Say, the α-level has not been reached yet, and we can add one more point: either $(4,0)$ or $(3,2)$. To maximize the probability to reject $H_{0T}\cap H_{0D}$ when $p_T=p_{AT}$, as in the LJL method, because of (2), we need to ensure good power when the probability of SD is 0 and hence $X_S=0$. Therefore point $(4,0)$ will be chosen. Alternatively, for good power for testing $H_{0T}\cap H_{0D}$, point $(3,2)$ will be chosen. If the sole focus is on testing H_{0T} or H_{0D}, the LJL test can be improved by considering a hybrid with the proposed approach where a rejection region is first constructed by selecting two linear boundaries as in the LJL test and then more points are added to the rejection region using the ordering described here in Section 2. This approach is more computationally intensive compared to the proposed method as multiple rejection regions exist for given α.

4. TESTING INDIVIDUAL HYPOTHESES H_{0T} AND H_{0D} IN A TWO-STAGE DESIGN

It is often of interest to test H_{0T} and H_{0D} as well as $H_{0T}\cap H_{0D}$. Thus far, we have considered only the type I error rate under $H_{0T}\cap H_{0D}$. To test also H_{0T} and H_{0D} when the trial continues to the maximum sample size, we need to find rejection regions for H_{0T} and $H_{0D}, \bar{S}_{2T}$and $\bar{S}_{2D}$ such that

$$
\begin{aligned}
&\Pr\big\{(X_T^{(1)},X_D^{(1)})\in\bar{S}_1 \text{ and}\\
&\qquad (X_T,X_D)\in\bar{S}_2 \mid p_{0T},p_{0D}\big\}\leq\alpha,\\
&\max_{p_D}\Pr\big\{(X_T^{(1)},X_D^{(1)})\in\bar{S}_1 \text{ and}\\
&\qquad\qquad (X_T,X_D)\in\bar{S}_{2T} \mid p_{0T},p_D\big\}\leq\alpha,\\
&\max_{p_T}\Pr\big\{(X_T^{(1)},X_D^{(1)})\in\bar{S}_1 \text{ and}\\
&\qquad\qquad (X_T,X_D)\in\bar{S}_{2D} \mid p_T,p_{0D}\big\}\leq\alpha,\\
&\Pr\big\{(X_T^{(1)},X_D^{(1)})\in\bar{S}_1 \text{ and}\\
&\qquad (X_T,X_D)\in\bar{S}_2 \mid p_{AT},p_{AD}\big\}\geq 1-\beta,\\
&\Pr\big\{(X_T^{(1)},X_D^{(1)})\in\bar{S}_1 \text{ and}\\
&\qquad (X_T,X_D)\in\bar{S}_{2T} \mid p_{AT},p_D=p_{AT}\big\}\geq 1-\beta_T,\\
&\Pr\big\{(X_T^{(1)},X_D^{(1)})\in\bar{S}_1 \text{ and}\\
&\qquad (X_T,X_D)\in\bar{S}_{2D} \mid p_T=0,p_{AD}\big\}\geq 1-\beta_D.
\end{aligned}
$$

Applying proposition A.1 from [12], the type I error rate to test H_{0T} is maximized when $p_D = 1 - p_T$. When $p_D = 1 - p_T$, the probability of stopping for futility is 0 and the rejection region coincides with the one for a single stage design with binary outcome. The type I error rate to test H_{0D} is maximized when $p_D = p_T$, that is, when $\Pr\{\mathrm{SD}\} = 0$. The probability of stopping for futility is then $\Pr\{(X_T^{(1)}, X_D^{(1)}) \in \bar{S}_1 \mid p_T = p_D = p_{0D}\}$. Conceivably, $\bar{S}_{2T}$ or $\bar{S}_{2D}$ can have points that are not in $\bar{S}_2$. When this occurs, to control type I error rate in a strong sense using the closed testing principle [14], the rejection region of H_{0T} is $\bar{S}_{2T} \cap \bar{S}_2$ and the rejection region of H_{0D} is $\bar{S}_{2D} \cap \bar{S}_2$.

Consider the example from Section 3. The rejection region for H_{0T}, is $\bar{S}_{2T} = \{X_T : X_T \geq 4\}$. Rejection region for H_{0D} is $\bar{S}_{2D}$ is found so that the probability to reject the null hypothesis $p_D = 0.35$ in a two-stage design with $n = 7$ and $n_1 = 5$ and stopping for futility after stage 1 if $X_{1D} \leq 1$ is at most α. We compute that $\bar{S}_{2D} = \{X_D : X_D \geq 6\}$. In this example, the power to reject $H_{0T} \cap H_{0D}$ when $p_T = p_{AT}$ is the same as the power to reject H_{0T}, and also the power to reject $H_{0T} \cap H_{0D}$ when $p_D = p_{AD}$ is the same as the power to reject H_{0D}. However in most cases, the power to reject H_{0T} will be smaller than the probability to reject $H_{0T} \cap H_{0D}$when $p_T = p_{AT}$, and similarly to H_{0D}. See our example in Section 5.

5. MOTIVATING EXAMPLE

Our motivating example is a single arm phase II trial of a novel agent that may have activity in non-small cell lung cancer in patients who have experienced disease progression after first-line therapy. We were interested in using disease control, defined as CR or PR or SD lasting more than 12 weeks, as the primary endpoint. The DC rates were set to $p_{0D} = 0.25$ and $p_{AD} = 0.50$. Tumor response was also of interest with $p_{0T} = 0.05$ and $p_{AT} = 0.25$. The Simon's minimax design to test H_{0D} alone with type I error rate of 0.05 and power of 0.80 yields the sample size of 24, H_{0T} alone can be tested with 16 patients. We tested $H_{0T} \cap H_{0D}$ with type I error rate of at most 0.05, and power to reject $H_{0T} \cap H_{0D}$ of 0.8 when $\{p_T = p_{AT}$ and $p_D = p_{AD}\}$, and at least 0.65 when $p_T = p_{AT}$ or $p_D = p_{AD}$. We obtained all possible designs using our software with parameters $p_{0T} = 0.05, p_{0D} = 0.25, p_{AT} = 0.25, p_{AD} = 0.5, \alpha = 0.05, \beta = 0.21, \beta_T = 0.35$ and $\beta_D = 0.35$. The minimax design to test $H_{0T} \cap H_{0D}$ was selected with $n = 18$ yielding the probability to reject $H_{0T} \cap H_{0D}$ of 0.80, 0.76 and 0.70 when $\{p_T = p_{AT}$ and $p_D = p_{AD}\}, p_T = p_{AT}$ and $p_D = p_{AD}$ respectively. The design has stage 1 size of $n_1 = 12$, futility region S_1 described by $t_1 = 1, d_1 = 5, A_1 = \{(2,2),(2,3),(2,4),(0,6)\}$ and S_2 by $t_2 = 2, d_2 = 7$ and $A_2 = \{\varnothing\}$. That is, $H_{0T} \cap H_{0D}$ is rejected at the end if $X_{1T} \geq 3$ or $X_{1D} \geq 8$. If one would like to test H_{0T} and H_{0D} as well as $H_{0T} \cap H_{0D}, H_{0T}$ is rejected if $X_{1T} \geq 4$, and H_{0D} is rejected if $X_{1T} \geq 9$.

6. DISCUSSION

Our motivation for this proposed method resulted from interest by the investigators' from the cancer center in testing $H_{0T} \cap H_{0D}$ rather than testing either H_{0T} or H_{0D}. Testing $H_{0T} \cap H_{0D}$ usually yields a smaller sample size than sample sizes for testing H_{0T} and H_{0D} separately. For example, recall that in our motivating example, the Simon's design to test H_{0D} alone with type I error rate of 0.05 and power of 0.8 yields the minimum sample size of 24. If we modify p_{AT} slightly and use $p_{AT} = 0.22$ in place of $p_{AT} = 0.25, 24$ patients are required to test H_{0T} alone. In comparison, if we test $H_{0T} \cap H_{0D}$ with 0.8 power, we need only 19 patients.

We have developed easy-to-use software to generate designs we describe here. From the user's input values of α, $1 - \beta$, $1 - \beta_T$, $1 - \beta_D$, p_{0T}, p_{0D}, p_{AT}, p_{AD}, the software calculates the sample sizes (n and n_1), the futility region points for stage 1 and final analysis, PET and EN_0, for each admissible design. The website also provides a suggested description of the method for a clinical trial protocol.

Apart from the proposed method, our software also contains various phase II methods that are frequently used at the Lineberger Comprehensive Cancer Center. Among methods available at http://cancer.unc.edu/biostatistics/program/ivanova/ are Simon's and Fleming's two-stage designs and the method to generate stopping boundary for continuous toxicity monitoring in a phase II trial [7].

ACKNOWLEDGEMENTS

Software development and Ivanova's work was supported in part by the National Institutes of Health RO1 CA120082-01A1. The authors thank Vladimir Khvostov for his expertise in developing software.

Received 5 August 2012

REFERENCES

[1] Abrey, L. E., Olson, J. D., Raizer, J. J., Mack, M., Rodavitch, A., Boutros, D. Y. and Malkin, M. G. (2001). A phase II trial of temozolomide for patients with recurrent or progressive brain metastases. *Journal of Neuro-Oncology* **53** 259–265.

[2] Cohen, E. E. W., Rosen, F., Stadler, W. M., Recant, W., Stenson, K., Huo, D. and Vokes, E. E. (2003). Phase II trial of ZD1839 in recurrent or metastatic squamous cell carcinoma of the head and neck. *Journal of Clinical Oncology* **21** 1980–1987.

[3] Eisenhauer, E. A, Therasse, P., Bogaerts, J., Schwartz, L. H., Sargent, D., Ford, R., Dancey, J., Arbuck, S., Gwyther, S., Mooney, M., Rubinstein, L., Shankar, L., Dodd, L., Kaplan, R., Lacombe, D. and Verweij, J. (2009). New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer* **45** 228–247.

[4] Erasmus, J. J., Gladish, G. W., Broemeling, L., Sabloff, B. S., Truong, M. T., Herbst, R. S. and Munden, R. F. (2003). Interobserver and intraobserver variability in measurement of non-small-cell carcinoma lung lesions: Implications for assessment of tumor response. *Journal of Clinical Oncology* **21** 2574–2582.

[5] Fleming, T. R. (1982). One-sample multiple testing procedure for phase II clinical trials. *Biometrics* **38** 143–151.

[6] Gutierrez, M. E., Kummar, S. and Giaccone, G. (2009). Next generation oncology drug development: Opportunities and challenges. *Nature Reviews Clinical Oncology* **6** 259–265.

[7] Ivanova, A., Qaqish, B. F. and Schell, M. J. (2005). Continuous toxicity monitoring in phase I trials in oncology. *Biometrics* **61** 540–545. MR2140926

[8] Jung, S. H., Lee, T., Kim, K. M. and George, S. L. (2004). Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* **23** 561–569.

[9] Kim, E. S., Hirsh, V., Mok, T., Socinski, M. A., Gervais, R., Wu, Y. L., Li, L. Y., Watkins, C. L., Sellers, M. V., Lowe, E. S., Sun, Y., Liao, M. L., Osterlind, K., Reck, M., Armour, A. A., Shepherd, F. A., Lippman, S. M. and Douillard, J. Y. (2008). Gefitinib Versus Docetaxel in previously treated non-small-cell lung cancer: A randomised phase III trial. *Lancet* **372** 1809–1818.

[10] Lin, X., Allred, R. and Andrews, G. (2008). A two-stage phase II trial design utilizing both primary and secondary endpoints. *Pharmaceutical Statistics* **7** 88–92.

[11] Lin, S. and Chen, T. (1998). Optimal two-stage designs for phase II trials with differentiation of complete and partial responses. *Communications in Statistics – Theory and Methods* **29** 923–940.

[12] Lu, Y., Jin, H. and Lamborn, K. R. (2005). A design of phase II cancer trials using total and complete response endpoints. *Statistics in Medicine* **24** 3155–3170. MR2209049

[13] Lynch, T. J., Bell, D. W., Sordella, R., Gurubhagavatula, S., Okimoto, R. A., Brannigan, B. W., Harris, P. L., Haserlat, S. M., Supko, J. G., Haluska, F. G., Louis, D. N., Christiani, D. C., Settleman, J. and Haber, D. A. (2004). Activating mutations in the epidermal growth factor receptor underlying responsiveness of non-small-cell lung cancer to Gefitinib. *New England Journal of Medicine* **350** 2129–2139.

[14] Marcus, R., Peritz, E. and Gabriel, K. R. (1976). On closed testing procedures with special reference to ordered analysis of variance. *Biometrika* **63** 655–660. MR0468056

[15] Oxnard, G. R., Zhao, B., Sima, C. S., Ginsberg, M. S., James, L. P., Lefkowitz, R. A., Guo, P., Kris, M. G., Schwartz, L. H. and Riely, G. J. (2011). Variability of Lung Tumor measurements on repeat computed tomography scans taken within 15 minutes. *Journal of Clinical Oncology* **29** 3114–3119.

[16] Paez, J. G., Janne, P. A., Lee, J. C., Tracy, S., Greulich, H., Gabriel, S., Herman, P., Kaye, F. J., Lindeman, N., Boggon, T. J., Naoki, K., Sasaki, H., Fujii, Y., Eck, Sellers, W. R., Johnson, B. E. and Meyerson, M. (2004). EGFR mutations in lung cancer: Correlation with clinical response to Gefitinib therapy. *Science* **304** 1497–1500.

[17] Panageas, K. (2002). An optimal two-stage phase II design utilizing complete and partial response information separately. *Controlled Clinical Trials* **23** 367–379.

[18] Shepherd, F. A., Dancey, J., Ramlau, R., Mattson, K., Gralla, R., O'Rourke, M., Levitan, N., Gressot, L., Vincent, M., Burkes, R., Coughlin, S., Kim, Y. and Berille, J. (2000). Prospective randomized trial of Docetaxel versus best supportive care in patients with non-small-cell lung cancer previously treated with platinum-based chemotherapy. *Journal Clinical Oncology* **18** 2095–2103.

[19] Shepherd, F. A., Pereira, J. R., Ciuleanu, T., Tan, E. H., Hirsh, V., Thongprasert, S., Campos, D., Maoleekoonpiroj, S., Smylie, M., Martins, R., van Kooten, M., Dediu, M., Findlay, B., Tu, D., Johnston, D., Bezjak, A., Clark, G., Santabarbara, P. and Seymour, L. (2005). Erlotinib in previously treated non-small-cell lung cancer. *New England Journal of Medicine* **353** 123–132.

[20] Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10** 1–10.

Anastasia Ivanova
Department of Biostatistics, CB #7420
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7420
USA
Lineberger Cancer Center at the University
of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27559-7305
USA
E-mail address: aivanova@bios.unc.edu

Jane Monaco
Department of Biostatistics, CB #7420
The University of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27599-7420
USA

Thomas Stinchcombe
Lineberger Cancer Center at the University
of North Carolina at Chapel Hill
Chapel Hill, North Carolina 27559-7305
USA