

Randomized phase II trials with a prospective control

Sin-Ho Jung^{*,†}

Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, U.S.A.

SUMMARY

We consider phase II trials randomizing patients between a prospective control and an experimental therapy. Proposed are two-stage designs allowing for early termination of the study when the experimental arm does not show promising efficacy at the interim analysis. By using exact binomial distributions, the design characteristics, such as type I error and power, are exactly calculated. Given response probabilities for two arms, we define minimax and optimal designs that satisfy a prespecified restriction on type I and II error probabilities. These designs are randomized phase II trial analogs of Simon's designs that were proposed for single-arm phase II trials. The methods for two-arm trials are easily extended to multi-arm trials with one control and $K (\geq 2)$ experimental arms. Some phase II trials are taken as real examples. Copyright © 2007 John Wiley & Sons, Ltd.

KEY WORDS: balanced allocation; futility; minimax design; multi-arm trial; optimal design; two-stage design

1. INTRODUCTION

Phase II clinical trials are to screen out experimental therapies with low tumor efficacy (tumor response) before a large-scale phase III trial is launched. Typically, a phase II study is conducted as a single-arm trial by accruing patients only to the experimental therapy to be compared with a historical control, e.g. Simon [1], Jung and Kim [2], Jung *et al.* [3]. The response rate of a historical control is usually estimated from a previous phase II trial. In practice, it is not possible to randomly select patients to join a trial from the entire patient population of interest. Therefore, the patient populations for different phase II trials are often quite heterogeneous, so that the distributions of patient characteristics for a new phase II trial are quite different from that of a prior phase II trial which provides the historical control. Furthermore, the response assessment may be different between the new study and the study of a historical control. In this case, the comparability of a new study with a historical control is questionable.

*Correspondence to: Sin-Ho Jung, Department of Biostatistics and Bioinformatics, Duke University, Durham, NC 27710, U.S.A.

†E-mail: sinho.jung@duke.edu, jung0005@mc.duke.edu

Like most phase II trials, a historical control study may have a small sample size, so that the estimated response rate will have a large variation. In designing a new single-arm phase II trial, however, we treat the response rate for the historical control like the true parameter value. Even if all the conditions are comparable between a historical control and a new single-arm trial, regarding the response rate for the historical control as a parameter can drastically increase the type I error in testing. Due to these reasons, some investigators feel more comfortable with a randomized phase II trial with a prospective control than with a single-arm trial to be compared to a historical control.

Study 50502 of Cancer and Leukemia Group B (CALGB) is a phase II trial to evaluate the tumor response of CD30 antibody, SGN-30, combined with GVD chemotherapy in patients with relapsed or refractory classical Hodgkin's lymphoma (HL). In a previous study (CALGB 59804), GVD has led to responses in 65 per cent (32/49) of patients with relapsed or refractory HL who never had a transplant and in 77 per cent (30/39) in the transplant group. Combining the data from the two cohorts, the new CALGB study (50502) was originally designed as a single-arm trial for testing:

$$H_0 : p \leq 70 \text{ per cent} \quad \text{against} \quad H_1 : p > 70 \text{ per cent}$$

where p denotes the true response probability for the combination therapy. Here, $p_0 = 70$ per cent is based on the estimated response probability from the pooled data of the previous study, i.e. $(32 + 30)/(49 + 39)$.

Some investigators are concerned that, although the combination therapy really has a larger response rate than the control therapy, the new single-arm trial may end up as a negative study if the new trial accrues more non-transplant patients. On the other hand, if the new trial accrues more transplant patients, the combination therapy may erroneously proceed to a large-scale phase III trial with false-positive findings and waste valuable resources. Because of these concerns, CALGB 50502 was later redesigned as a randomized phase II trial with a GVD plus placebo arm as a prospective control. The randomization is stratified by transplant (yes/no).

There are some published statistical approaches for designing and analyzing randomized phase II trials. Simon *et al.* [4] propose the play-the-winner selection rule for randomized phase II trials with experimental arms only. Liu *et al.* [5] review Simon *et al.* [4] and point out that this approach has a high selection probability even when the treatment arms have an equal efficacy. Sargent and Goldberg [6] consider a similar approach to Simon *et al.* [4] by allowing for selection based on other factors in case the difference in observed response rates is small.

Thall *et al.* [7] propose a two-stage screening procedure with a control and K experimental arms. In the first stage, they randomize n_1 patients to each of the K experimental arms, and pick the winner for the second stage if its observed efficacy is larger than that for the historical control by 10 per cent (and stop the trial otherwise). In the second stage, they randomize n_2 patients into each of the control arm and the winner experimental arm from stage 1, and conduct a one-sided chi-square test to see if the experimental arm is better than the control.

Palmer [8] proposes a two-stage design for a three-arm selection case. In stage 1, a cohort of three patients is randomized to arms A, B and C, and he decides to either continue to accrue the next cohort or stop stage 1 after choosing the best two arms. In stage 2, a cohort of two patients is randomized to the two arms chosen at stage 1, and he decides to either continue to accrue the next cohort or stop stage 2 by choosing the winner. Given the maximum number of patients available for a study, the stopping time for each stage is chosen to minimize the number of future failures using a Bayesian approach. This method requires quick responses for the sequential tests.

Steinberg and Venzon [9] propose a two-stage design for a phase II trial with two experimental arms. In stage 1, n_1 patients are randomized to each arm. The trial is stopped after stage 1 if the difference in number of responders between the two arms is larger than d , which is chosen so that, when the two arms have a 15 per cent difference in response rates, the probability of selecting the inferior arm is controlled at a specified level. Otherwise, the trial proceeds to stage 2 to randomize n_2 patients to each arm. After stage 2, they pick the winner based on the cumulative responses through the two stages. Given $n = n_1 + n_2$, they propose to choose $n_1 = n_2 = n/2$ or to minimize the expected sample size for the specified response rates with a 15 per cent difference. This approach does not control the overall error probabilities through two stages.

Rubinstein *et al.* [10] discuss the strengths and weaknesses of some existing design methods for randomized phase II trials, and propose a method for randomized phase II screening designs based on large-sample approximation.

Most of these publications on selection trials consider randomizing patients to experimental arms only, and selecting efficacious experimental therapies to be investigated further. Screening trial designs for selecting among experimental treatments should not be used for comparing experimental treatments with a control. Furthermore, most of these methods for randomized phase II trials have common drawbacks. They do not accurately control the type I error and power reflected by the small sample sizes and multi-stage designs of randomized phase II trials.

In this paper, while addressing these issues, we develop efficient design and testing methods for randomized phase II trials with a prospective control. Our methods are to test whether an experimental arm has a higher response probability than a prospective control. The two-stage randomized phase II trial designs allow for early stopping for futility of the experimental arm. We propose Simon-type [1] minimax and optimal designs for randomized phase II trials. The proposed methods can be easily extended to the cases of randomizing patients to a control and $K (\geq 2)$ experimental arms. Some CALGB studies including the aforementioned 50502 are taken as real examples.

2. TWO-STAGE DESIGNS

In this section, we consider two-stage designs for randomized phase II trials between an experimental arm (Arm 1) and a control arm (Arm 2). Trials with two or more experimental arms and a prospective control will be investigated in the following section. Let p_1 and p_2 denote the true response rates for Arms 1 and 2, respectively. We want to test whether the experimental arm has a higher response rate than the control or not, i.e. $H_0 : p_1 \leq p_2$ against $H_1 : p_1 > p_2$.

In the first stage, we accrue n_1 patients to each arm. Let X_1 and Y_1 denote the number of responders among the n_1 first stage patients for Arms 1 and 2, respectively. We proceed to the second stage if $X_1 - Y_1 \geq a_1$ for a chosen integer $a_1 \in [-n_1, n_1]$. Otherwise, we reject Arm 1 (or fail to reject H_0) and stop the trial. In the second stage, we accrue an additional n_2 patients to each arm. Let X_2 and Y_2 denote the number of responders among the second stage patients of Arms 1 and 2, respectively. Also, let $X = X_1 + X_2$ and $Y = Y_1 + Y_2$ denote the total number of responders among the cumulative $n = n_1 + n_2$ patients for Arms 1 and 2, respectively. For an integer $a \in [a_1 - n_2, n]$, we accept Arm 1 (or reject H_0) for further investigation if $X - Y \geq a$. Otherwise, we reject Arm 1.

Let p_0 denote the projected response rate for the historical control and Δ a clinically significant increase in response rate for Arm 1. For the purpose of type I and II error calculations, we specify a

point null hypothesis $H_0 : p_1 = p_2 = p_0$ and an alternative hypothesis $H_1 : (p_1, p_2) = (p_0 + \Delta, p_0)$. For a two-stage design defined by (n_1, n, a_1, a) , the type I error and power of the two-stage design are calculated as

$$\alpha = P(X_1 - Y_1 \geq a_1, X - Y \geq a | p_1 = p_2 = p_0)$$

and

$$1 - \beta = P(X_1 - Y_1 \geq a_1, X - Y \geq a | p_1 = p_0 + \Delta, p_2 = p_0)$$

respectively. Let $B(n, p)$ denote the binomial distribution with n independent trials and a probability of success p for each trial. These probabilities are calculated assuming that $X_1 \sim B(n_1, p_1)$, $X_2 \sim B(n_2, p_1)$, $Y_1 \sim B(n_1, p_2)$ and $Y_2 \sim B(n_2, p_2)$ are independent random variables. That is,

$$\begin{aligned} \alpha = & \sum_{k_1=a_1}^{n_1} \sum_{y_1=\max(0, -k_1)}^{n_1-\max(0, k_1)} \sum_{k_2=a-k_1}^{n_2} \sum_{y_2=\max(0, -k_2)}^{n_2-\max(0, k_2)} b(y_1|n_1, p_0)b(k_1 + y_1|n_1, p_0) \\ & \times b(y_2|n_2, p_0)b(k_2 + y_2|n_2, p_0) \end{aligned}$$

and

$$\begin{aligned} 1 - \beta = & \sum_{k_1=a_1}^{n_1} \sum_{y_1=\max(0, -k_1)}^{n_1-\max(0, k_1)} \sum_{k_2=a-k_1}^{n_2} \sum_{y_2=\max(0, -k_2)}^{n_2-\max(0, k_2)} b(y_1|n_1, p_0)b(k_1 + y_1|n_1, p_1) \\ & \times b(y_2|n_2, p_0)b(k_2 + y_2|n_2, p_1) \end{aligned}$$

where $b(x|n, p) = \binom{n}{x} p^x (1-p)^{n-x}$ for $x = 0, 1, \dots, n$ denotes the probability mass function of the binomial distribution with n trials and a probability of success p .

Suppose that we want to choose a two-stage design with type I error smaller than or equal to α^* and power larger than or equal to $1 - \beta^*$. There exist many two-stage designs satisfying the (α^*, β^*) -restriction. We next define two reasonable two-stage designs for a randomized phase II trial.

2.1. Optimal design

We wish to find the two-stage design with the smallest expected sample size when the experimental therapy has a low response rate, i.e. as specified under H_0 . The probability of early termination under $H_0 : p_1 = p_2 = p_0$ is calculated as

$$\text{PET} = P(X_1 - Y_1 < a_1 | p_0) = \sum_{k_1=-n_1}^{a_1-1} \sum_{y_1=\max(0, -k_1)}^{n_1-\max(0, k_1)} b(y_1|n_1, p_0)b(k_1 + y_1|n_1, p_0)$$

Since, under H_0 , the sample size per arm is n_1 with probability PET and n with probability $1 - \text{PET}$, the expected sample size per arm under H_0 is obtained as

$$\text{EN} = n_1 \times \text{PET} + n(1 - \text{PET})$$

Among the two-stage designs satisfying the (α^*, β^*) -restriction, the ‘optimal design’ is defined as the one with the smallest EN.

2.2. Minimax design

Among the two-stage designs satisfying the (α^*, β^*) -restriction, the ‘minimax design’ is defined as the one with the smallest maximal sample size n . For the chosen n , there may be more than one two-stage designs satisfying the (α^*, β^*) -restriction. In this case, we choose the design with the smallest EN as the minimax design.

Given n , the designs satisfying the (α^*, β^*) -restriction can be determined by an exhaustive enumeration by changing n_1 , a_1 and a ($1 \leq n_1 \leq n-1$, $-n_1 \leq a_1 \leq n_1$, $a_1 - n_2 \leq a \leq n$). Among these designs, the one that minimizes EN is identified. For the given n , this design dominates all other designs in terms of EN. Let $D(n) = (n_1, n, a_1, a)$ denote the design with the smallest EN, $EN(n)$, among the designs with maximal sample size n satisfying the (α^*, β^*) -restriction. If n is too small, there may exist no designs satisfying the (α^*, β^*) -restriction. The design $D(n)$ with the smallest n is the minimax design. If n exceeds a certain limit, the two-stage design practically becomes identical to the single-stage design with the minimal sample size in the sense that the critical value of the first stage of the two-stage design is the same as that of the corresponding single-stage design and no decision is made in the second stage. Hence, as n increases beyond the limit, $EN(n)$ increases linearly. The search for the optimal design continues by checking $EN(n)$ until n becomes so large that $EN(n)$ starts to linearly increase in n .

Tables I–IV report the minimax and optimal designs under $\alpha^* = 0.15, 0.2$; $1 - \beta^* = 0.8, 0.85$; $p_0 = 0.05 : 0.85(0.05)$; $\Delta = 0.15, 0.2$ (0.1 also for $p = 0.05$ and 0.85). We also list the single-stage design (n, a) under each setting of $(p_0, p_1, \alpha^*, 1 - \beta^*)$, where a is the critical value to reject H_0 if the number of responders among n patients in the experimental arm minus that in the control arm is larger than or equal to a . Note that, under each setting, the maximal sample size for the minimax design is smaller than or equal to the sample size of the single-stage design. Under some settings, the single-stage design requires more patients than the maximal sample size of the optimal design.

2.3. An example

An example is taken from the CALGB study that was briefly discussed in the Introduction.

Example 1

CALGB 50502 is a randomized phase II trial to evaluate the anti-tumor activity of CD30 antibody, SGN-30, combined with GVD chemotherapy (Arm 1) compared with GVD plus placebo (Arm 2) in patients with relapsed/refractory classical HL. The randomization is stratified by prior stem cell transplant (yes/no). The primary objective of this study is for testing $H_0 : p_1 \leq p_2$ against $H_1 : p_1 > p_2$. For the purpose of type I error and power calculation, the hypotheses are specified as $H_0 : p_1 = p_2 = 0.7$ and $H_1 : p_1 = 0.85, p_2 = 0.7$, i.e. $p_0 = 0.7$ and $\Delta = 0.15$. The design parameters are chosen based on the historical data (CALGB 59804) and a clinically significant difference determined by study investigators. Under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.7, 0.85, 0.15, 0.8)$, the minimax design is $(n_1, n, a_1, a) = (31, 63, -1, 6)$, which has $(\alpha, 1 - \beta, EN) = (0.1392, 0.8002, 52.16)$, and the optimal design is $(n_1, n, a_1, a) = (27, 73, 1, 6)$ which has $(\alpha, 1 - \beta, EN) = (0.1321, 0.8001, 47.28)$. Compared to the minimax design, the optimal requires 10 more patients per arm in the maximal sample size, but saves almost five patients per arm in expected sample size when Arm 1 is inefficient. Considering the large difference in the maximal sample size, we may choose the minimax design for the trial. The single-stage design under the same design parameters requires $n = 63$ patients to reject H_0 when $X - Y \geq 6$, for which $(\alpha, 1 - \beta) = (0.1423, 0.8046)$. Note that this sample size is equal to the maximal sample size for the minimax two-stage design, but, compared

Table I. Single-stage designs, and minimax and optimal two-stage designs for $(\alpha^*, 1 - \beta^*) = (0.15, 0.8)$ and balanced allocation ($r = 1$).

P_0	P_1	Single-stage design			Minimax design			Optimal design				
		(n, a)	α	$1 - \beta$	(n, n_1, a_1, a)	α	$1 - \beta$	EN	(n, n_1, a_1, a)	α	$1 - \beta$	EN
		0.05	0.15	(50, 3)	0.1211	0.8018	(50, 29, -1, 3)	0.1209	0.8000	46.37	(57, 20, 5, 9)	0.1208
	0.2	(22, 2)	0.1381	0.8016	(22, 12, -1, 2)	0.1380	0.8006	21.29	(27, 16, 1, 2)	0.1395	0.8015	19.55
	0.25	(16, 2)	0.1003	0.8113	(16, 8, 0, 2)	0.0995	0.8001	14.08	(22, 10, 1, 2)	0.1028	0.8034	13.22
0.1	0.25	(43, 4)	0.1022	0.8041	(38, 30, 2, 3)	0.1479	0.8022	32.04	(56, 17, 1, 3)	0.1499	0.8007	31.94
	0.3	(24, 3)	0.1106	0.8054	(24, 15, 0, 3)	0.1099	0.8002	20.62	(28, 14, 1, 3)	0.1118	0.8006	19.18
0.15	0.3	(54, 5)	0.1117	0.8010	(46, 36, 1, 4)	0.1500	0.8006	40.34	(51, 28, 1, 4)	0.1479	0.8004	37.78
	0.35	(32, 4)	0.1086	0.8060	(32, 21, 2, 3)	0.1461	0.8025	23.81	(32, 21, 2, 3)	0.1461	0.8025	23.81
0.2	0.35	(57, 5)	0.1454	0.8060	(57, 38, 0, 5)	0.1436	0.8006	48.59	(66, 30, 1, 5)	0.1407	0.8006	45.68
	0.4	(33, 4)	0.1397	0.8040	(33, 23, 0, 4)	0.1388	0.8009	28.74	(39, 18, 1, 4)	0.1369	0.8024	26.75
0.25	0.4	(67, 6)	0.1360	0.8027	(67, 42, -1, 6)	0.1351	0.8001	58.19	(78, 33, 1, 6)	0.1324	0.8004	52.95
	0.45	(41, 5)	0.1250	0.8100	(39, 31, 3, 4)	0.1499	0.8046	32.85	(43, 24, 2, 4)	0.1486	0.8013	29.85
0.3	0.45	(77, 7)	0.1263	0.8029	(70, 48, 1, 6)	0.1499	0.8008	58.02	(77, 37, 1, 6)	0.1468	0.8002	54.98
	0.5	(41, 5)	0.1388	0.8034	(41, 28, 0, 5)	0.1376	0.8002	35.25	(46, 22, 1, 5)	0.1345	0.8004	32.43
0.35	0.5	(78, 7)	0.1375	0.8032	(78, 48, -1, 7)	0.1362	0.8000	66.78	(95, 39, 2, 6)	0.1497	0.8006	59.21
	0.55	(48, 6)	0.1194	0.8052	(42, 25, 0, 5)	0.1477	0.8004	34.50	(50, 20, 1, 5)	0.1429	0.8001	33.03
0.4	0.55	(78, 7)	0.1440	0.8008	(78, 57, -1, 7)	0.1437	0.8001	69.87	(90, 38, 1, 7)	0.1387	0.8001	61.58
	0.6	(48, 6)	0.1258	0.8040	(45, 30, 2, 5)	0.1491	0.8011	35.20	(49, 21, 1, 5)	0.1489	0.8010	33.25
0.45	0.6	(78, 7)	0.1477	0.8008	(78, 57, -1, 7)	0.1474	0.8001	69.83	(90, 38, 1, 7)	0.1416	0.8001	61.62
	0.65	(48, 6)	0.1295	0.8052	(46, 34, 3, 5)	0.1483	0.8013	37.26	(49, 27, 2, 5)	0.1476	0.8007	34.50
0.5	0.65	(78, 7)	0.1490	0.8032	(78, 48, -1, 7)	0.1472	0.8000	66.61	(90, 37, 1, 7)	0.1476	0.8003	61.05
	0.7	(48, 6)	0.1307	0.8090	(45, 34, 3, 5)	0.1493	0.8026	37.00	(53, 18, 1, 5)	0.1497	0.8011	33.19
0.55	0.7	(77, 7)	0.1462	0.8029	(77, 48, -1, 7)	0.1447	0.8003	66.00	(90, 35, 1, 7)	0.1386	0.8007	59.87
	0.75	(47, 6)	0.1270	0.8063	(43, 29, 2, 5)	0.1485	0.8005	33.85	(48, 19, 1, 5)	0.1465	0.8001	31.63
0.6	0.75	(76, 7)	0.1409	0.8053	(76, 41, -1, 7)	0.1382	0.8006	63.13	(89, 37, 2, 6)	0.1496	0.8004	55.77
	0.8	(46, 6)	0.1208	0.8064	(40, 22, 0, 5)	0.1469	0.8012	32.10	(45, 19, 1, 5)	0.1404	0.8002	30.29
0.65	0.8	(74, 7)	0.1312	0.8052	(68, 47, 2, 6)	0.1498	0.8019	54.83	(75, 31, 1, 6)	0.1460	0.8004	50.67
	0.85	(38, 5)	0.1394	0.8014	(38, 26, 0, 5)	0.1382	0.8001	32.69	(44, 17, 1, 5)	0.1300	0.8007	28.57
0.7	0.85	(63, 6)	0.1423	0.8046	(63, 31, -1, 6)	0.1392	0.8002	52.16	(73, 27, 1, 6)	0.1321	0.8001	47.28
	0.9	(37, 5)	0.1265	0.8087	(33, 22, 2, 4)	0.1490	0.8003	25.42	(36, 14, 1, 4)	0.1475	0.8003	23.20
0.75	0.9	(60, 6)	0.1228	0.8049	(52, 28, 0, 5)	0.1477	0.8005	41.48	(60, 23, 1, 5)	0.1400	0.8021	38.99
	0.95	(29, 4)	0.1436	0.8115	(29, 11, 0, 4)	0.1340	0.8003	21.76	(33, 12, 1, 4)	0.1268	0.8008	20.53
0.8	0.95	(48, 5)	0.1248	0.8042	(42, 22, 1, 4)	0.1449	0.8013	30.49	(46, 18, 1, 4)	0.1449	0.8014	29.67
0.85	0.95	(89, 6)	0.1235	0.8058	(76, 48, 0, 5)	0.1495	0.8003	63.60	(89, 32, 1, 5)	0.1379	0.8003	56.50

Table II. Single-stage designs, and minimax and optimal two-stage designs for $(\alpha^*, 1 - \beta^*) = (0.15, 0.85)$ and balanced allocation ($r = 1$).

p_0	p_1	Single-stage design			Minimax design			Optimal design				
		(n, a)	α	$1 - \beta$	(n, n_1, a_1, a)	α	$1 - \beta$	EN	(n, n_1, a_1, a)	α	$1 - \beta$	EN
0.05	0.15	(58, 3)	0.1391	0.8518	(58, 36, -1, 3)	0.1389	0.8502	53.59	(69, 37, 1, 3)	0.1399	0.8502	49.49
	0.2	(36, 3)	0.0837	0.8581	(34, 27, 2, 2)	0.1386	0.8511	28.15	(34, 27, 2, 2)	0.1386	0.8511	28.15
	0.25	(19, 2)	0.1201	0.8653	(19, 15, 1, 2)	0.1158	0.8512	16.26	(22, 13, 1, 2)	0.1179	0.8507	15.70
0.1	0.25	(49, 4)	0.1173	0.8527	(49, 35, 0, 4)	0.1169	0.8505	43.12	(59, 32, 2, 3)	0.1499	0.8500	39.09
	0.3	(28, 3)	0.1291	0.8589	(28, 17, 0, 3)	0.1276	0.8516	23.78	(32, 17, 1, 3)	0.1276	0.8500	22.75
0.15	0.3	(62, 5)	0.1280	0.8534	(62, 43, 0, 5)	0.1271	0.8503	53.65	(73, 32, 1, 5)	0.1271	0.8503	49.63
	0.35	(36, 4)	0.1225	0.8502	(36, 27, -1, 4)	0.1225	0.8501	33.46	(44, 19, 1, 4)	0.1238	0.8502	29.22
0.2	0.35	(74, 6)	0.1287	0.8523	(74, 53, 0, 6)	0.1280	0.8502	64.52	(89, 41, 2, 5)	0.1493	0.8506	57.26
	0.4	(45, 5)	0.1171	0.8557	(43, 29, 2, 4)	0.1487	0.8507	33.34	(45, 28, 2, 4)	0.1466	0.8501	33.22
0.25	0.4	(85, 7)	0.1246	0.8502	(78, 61, 2, 6)	0.1497	0.8502	67.40	(87, 42, 1, 6)	0.1477	0.8504	62.24
	0.45	(46, 5)	0.1388	0.8533	(46, 32, 0, 5)	0.1378	0.8506	39.81	(52, 25, 1, 5)	0.1354	0.8500	36.74
0.3	0.45	(87, 7)	0.1410	0.8504	(87, 66, -1, 7)	0.1408	0.8500	78.86	(98, 46, 1, 7)	0.1377	0.8502	69.64
	0.5	(54, 6)	0.1238	0.8561	(50, 35, 2, 5)	0.1492	0.8505	40.22	(55, 25, 1, 5)	0.1495	0.8510	38.16
0.35	0.5	(98, 8)	0.1306	0.8537	(91, 67, 2, 7)	0.1498	0.8502	76.43	(103, 45, 1, 7)	0.1475	0.8505	71.45
	0.55	(54, 6)	0.1334	0.8522	(54, 33, -1, 6)	0.1325	0.8501	46.66	(63, 26, 1, 6)	0.1298	0.8503	42.36
0.4	0.55	(98, 8)	0.1370	0.8512	(98, 68, -1, 8)	0.1364	0.8500	86.11	(112, 51, 2, 7)	0.1498	0.8504	74.24
	0.6	(54, 6)	0.1399	0.8510	(54, 37, -1, 6)	0.1395	0.8502	47.86	(61, 28, 1, 6)	0.1356	0.8507	42.71
0.45	0.6	(98, 8)	0.1408	0.8512	(98, 68, -1, 8)	0.1401	0.8500	86.06	(109, 61, 3, 7)	0.1499	0.8503	76.58
	0.65	(54, 6)	0.1437	0.8522	(54, 33, -1, 6)	0.1425	0.8501	46.54	(63, 26, 1, 6)	0.1378	0.8503	42.45
0.5	0.65	(98, 8)	0.1420	0.8537	(98, 59, -1, 8)	0.1400	0.8503	82.74	(100, 56, 0, 8)	0.1389	0.8508	79.65
	0.7	(54, 6)	0.1449	0.8561	(54, 33, 0, 6)	0.1416	0.8506	44.53	(60, 27, 1, 6)	0.1373	0.8502	41.72
0.55	0.7	(96, 8)	0.1383	0.8511	(96, 59, -2, 8)	0.1375	0.8500	84.09	(107, 50, 2, 7)	0.1499	0.8512	71.75
	0.75	(53, 6)	0.1414	0.8559	(53, 32, 0, 6)	0.1381	0.8508	43.55	(60, 25, 1, 6)	0.1336	0.8507	40.53
0.6	0.75	(94, 8)	0.1320	0.8511	(88, 62, 2, 7)	0.1500	0.8503	72.18	(99, 41, 1, 7)	0.1468	0.8506	67.40
	0.8	(51, 6)	0.1331	0.8511	(51, 31, -1, 6)	0.1321	0.8500	44.02	(55, 28, 2, 5)	0.1499	0.8520	37.22
0.65	0.8	(83, 7)	0.1450	0.8531	(83, 48, -1, 7)	0.1428	0.8502	69.91	(94, 39, 1, 7)	0.1371	0.8506	63.90
	0.85	(50, 6)	0.1243	0.8576	(44, 22, 0, 5)	0.1494	0.8513	34.38	(49, 21, 1, 5)	0.1439	0.8514	33.20
0.7	0.85	(79, 7)	0.1294	0.8502	(73, 58, 3, 6)	0.1498	0.8516	62.59	(82, 33, 1, 6)	0.1458	0.8503	54.88
	0.9	(41, 5)	0.1388	0.8551	(41, 22, 0, 5)	0.1342	0.8501	32.74	(47, 18, 1, 5)	0.1283	0.8525	30.40
0.75	0.9	(67, 6)	0.1360	0.8551	(67, 45, 3, 5)	0.1479	0.8511	50.96	(69, 36, 2, 5)	0.1494	0.8510	47.26
	0.95	(38, 5)	0.1161	0.8517	(32, 15, 0, 4)	0.1479	0.8504	24.93	(37, 14, 1, 4)	0.1385	0.8505	23.50
0.8	0.95	(53, 5)	0.1366	0.8524	(53, 31, 0, 5)	0.1334	0.8503	43.39	(60, 28, 2, 4)	0.1495	0.8510	37.83
0.85	0.95	(98, 6)	0.1350	0.8516	(95, 67, 3, 5)	0.1490	0.8513	74.61	(103, 51, 2, 5)	0.1469	0.8503	68.57

Table III. Single-stage designs, and minimax and optimal two-stage designs for $(\alpha^*, 1 - \beta^*) = (0.2, 0.8)$ and balanced allocation ($r = 1$).

P_0	P_1	Single-stage design			Minimax design			Optimal design				
		(n, a)	α	$1 - \beta$	(n, n_1, a_1, a)	α	$1 - \beta$	EN	(n, n_1, a_1, a)	α	$1 - \beta$	EN
0.05	0.15	(50, 3)	0.1211	0.8018	(41, 20, 4, 8)	0.1679	0.8032	30.20	(42, 20, 5, 8)	0.1677	0.8008	28.72
	0.2	(22, 2)	0.1381	0.8016	(22, 12, -1, 2)	0.1380	0.8006	21.29	(27, 16, 1, 2)	0.1395	0.8015	19.55
	0.25	(16, 2)	0.1003	0.8113	(16, 8, 0, 2)	0.0995	0.8001	14.08	(18, 9, 1, 1)	0.2000	0.8046	11.29
0.1	0.25	(34, 3)	0.1529	0.8023	(34, 26, 0, 3)	0.1524	0.8004	30.75	(35, 27, 2, 2)	0.1980	0.8073	28.95
	0.3	(17, 2)	0.1883	0.8015	(17, 10, -1, 2)	0.1882	0.8004	16.15	(18, 10, 0, 2)	0.1912	0.8027	15.25
0.15	0.3	(46, 4)	0.1521	0.8065	(41, 27, 1, 3)	0.1591	0.8031	32.93	(42, 26, 1, 3)	0.1972	0.8017	32.75
	0.35	(25, 3)	0.1586	0.8003	(25, 18, -1, 3)	0.1586	0.8000	23.33	(28, 17, 1, 3)	0.1566	0.8002	21.44
0.2	0.35	(48, 4)	0.1852	0.8050	(48, 29, -1, 4)	0.1835	0.8005	42.10	(53, 23, 0, 4)	0.1811	0.8002	40.21
	0.4	(27, 3)	0.1962	0.8120	(27, 17, 0, 3)	0.1918	0.8013	22.86	(29, 13, 0, 3)	0.1912	0.8021	22.57
0.25	0.4	(58, 5)	0.1670	0.8009	(53, 42, 2, 4)	0.1998	0.8016	45.88	(56, 33, 1, 4)	0.1980	0.8009	43.20
	0.45	(34, 4)	0.1629	0.8052	(32, 24, 2, 3)	0.1982	0.8022	26.46	(35, 17, 1, 3)	0.1992	0.8001	24.58
0.3	0.45	(60, 5)	0.1848	0.8051	(60, 37, -1, 5)	0.1825	0.8002	51.91	(68, 38, 2, 4)	0.1976	0.8010	48.61
	0.5	(35, 4)	0.1804	0.8090	(35, 23, 0, 4)	0.1769	0.8019	29.77	(41, 23, 2, 3)	0.1947	0.8013	28.66
0.35	0.5	(60, 5)	0.1945	0.8001	(60, 48, -2, 5)	0.1944	0.8000	56.44	(67, 30, 0, 5)	0.1886	0.8000	50.49
	0.55	(35, 4)	0.1901	0.8052	(35, 20, -1, 4)	0.1878	0.8005	30.36	(38, 17, 0, 4)	0.1843	0.8012	29.00
0.4	0.55	(70, 6)	0.1713	0.8043	(62, 35, -1, 5)	0.1996	0.8001	52.35	(69, 29, 0, 5)	0.1942	0.8004	51.13
	0.6	(35, 4)	0.1966	0.8040	(35, 21, -1, 4)	0.1944	0.8002	30.54	(39, 16, 0, 4)	0.1895	0.8013	29.14
0.45	0.6	(70, 6)	0.1751	0.8043	(64, 50, 2, 5)	0.1995	0.8024	55.34	(69, 29, 0, 5)	0.1969	0.8004	51.10
	0.65	(42, 5)	0.1618	0.8092	(35, 20, -1, 4)	0.1973	0.8005	30.24	(38, 17, 0, 4)	0.1924	0.8012	28.93
0.5	0.65	(69, 6)	0.1746	0.8012	(64, 35, 0, 5)	0.1998	0.8006	50.88	(67, 30, 0, 5)	0.1973	0.8000	50.40
	0.7	(41, 5)	0.1601	0.8034	(35, 17, -1, 4)	0.1964	0.8010	29.53	(38, 16, 0, 4)	0.1916	0.8021	28.54
0.55	0.7	(68, 6)	0.1716	0.8006	(61, 32, -1, 5)	0.1998	0.8009	50.76	(66, 29, 0, 5)	0.1944	0.8008	49.44
	0.75	(34, 4)	0.1969	0.8052	(34, 18, -1, 4)	0.1932	0.8002	29.07	(37, 15, 0, 4)	0.1872	0.8002	27.60
0.6	0.75	(58, 5)	0.1969	0.8009	(58, 38, -2, 5)	0.1961	0.8001	52.42	(63, 29, 0, 5)	0.1887	0.8003	47.81
	0.8	(33, 4)	0.1896	0.8040	(33, 18, -1, 4)	0.1868	0.8007	28.42	(36, 14, 0, 4)	0.1807	0.8023	26.68
0.65	0.8	(57, 5)	0.1884	0.8060	(57, 31, -1, 5)	0.1841	0.8004	48.03	(69, 24, 1, 4)	0.1995	0.8014	43.79
	0.85	(32, 4)	0.1794	0.8060	(32, 21, 2, 3)	0.1988	0.8025	24.45	(38, 13, 1, 3)	0.1981	0.8005	23.47
0.7	0.85	(54, 5)	0.1722	0.8010	(49, 30, 1, 4)	0.1987	0.8001	38.44	(51, 19, 0, 4)	0.1984	0.8005	37.25
	0.9	(31, 4)	0.1657	0.8129	(26, 16, 1, 3)	0.1997	0.8040	20.23	(28, 6, 0, 3)	0.1976	0.8016	19.73
0.75	0.9	(43, 4)	0.1913	0.8041	(43, 28, 0, 4)	0.1867	0.8001	36.42	(47, 16, 0, 4)	0.1789	0.8014	34.02
	0.95	(23, 3)	0.1965	0.8180	(23, 9, 0, 3)	0.1815	0.8035	17.51	(24, 5, 0, 3)	0.1756	0.8019	17.26
0.8	0.95	(40, 4)	0.1631	0.8079	(33, 12, 0, 3)	0.1852	0.8031	24.64	(34, 9, 0, 3)	0.1944	0.8008	24.45
0.85	0.95	(63, 4)	0.1903	0.8009	(63, 39, -1, 4)	0.1886	0.8001	55.42	(69, 40, 2, 3)	0.1994	0.8006	49.23

Table IV. Single-stage designs, and minimax and optimal two-stage designs for $(\alpha^*, 1 - \beta^*) = (0.2, 0.85)$ and balanced allocation ($r = 1$).

p_0	p_1	Single-stage design			Minimax design			Optimal design				
		(n, a)	α	$1 - \beta$	(n, n_1, a_1, a)	α	$1 - \beta$	EN	(n, n_1, a_1, a)	α	$1 - \beta$	EN
		0.05	0.15	(58, 3)	0.1391	0.8518	(58, 36, -1, 3)	0.1389	0.8502	53.59	(70, 20, 3, 8)	0.1833
	0.2	(26, 2)	0.1591	0.8523	(26, 20, 0, 2)	0.1588	0.8502	23.93	(27, 16, 0, 2)	0.1618	0.8516	23.45
	0.25	(19, 2)	0.1201	0.8653	(19, 15, 1, 2)	0.1158	0.8512	16.26	(22, 13, 1, 2)	0.1179	0.8507	15.70
0.1	0.25	(40, 3)	0.1729	0.8544	(40, 29, 0, 3)	0.1717	0.8505	35.47	(43, 21, 0, 3)	0.1732	0.8509	34.30
	0.3	(28, 3)	0.1291	0.8589	(27, 15, 1, 2)	0.1981	0.8500	19.50	(27, 15, 1, 2)	0.1981	0.8500	19.50
0.15	0.3	(53, 4)	0.1694	0.8541	(53, 38, 0, 4)	0.1678	0.8501	46.46	(59, 37, 2, 3)	0.1955	0.8508	43.85
	0.35	(30, 3)	0.1810	0.8589	(30, 20, 0, 3)	0.1785	0.8521	25.89	(32, 15, 0, 3)	0.1787	0.8512	25.25
0.2	0.35	(65, 5)	0.1614	0.8524	(58, 43, 1, 4)	0.1993	0.8508	49.69	(61, 31, 0, 4)	0.1990	0.8509	47.90
	0.4	(38, 4)	0.1569	0.8537	(36, 28, 2, 3)	0.1990	0.8557	30.46	(40, 20, 1, 3)	0.1988	0.8523	28.42
0.25	0.4	(67, 5)	0.1844	0.8502	(67, 51, -2, 5)	0.1843	0.8500	62.46	(74, 35, 0, 5)	0.1811	0.8505	56.64
	0.45	(39, 4)	0.1795	0.8518	(39, 26, -1, 4)	0.1788	0.8501	34.91	(43, 19, 0, 4)	0.1769	0.8504	32.79
0.3	0.45	(78, 6)	0.1682	0.8503	(70, 44, -1, 5)	0.1999	0.8501	60.55	(78, 35, 0, 5)	0.1951	0.8500	58.73
	0.5	(40, 4)	0.1963	0.8529	(40, 26, -1, 4)	0.1949	0.8503	35.45	(45, 19, 0, 4)	0.1909	0.8504	33.83
0.35	0.5	(80, 6)	0.1809	0.8535	(78, 62, 3, 5)	0.1998	0.8520	67.10	(85, 42, 1, 5)	0.1993	0.8500	61.54
	0.55	(47, 5)	0.1651	0.8501	(43, 34, 2, 4)	0.1983	0.8504	37.16	(45, 27, 1, 4)	0.1971	0.8500	34.98
0.4	0.55	(80, 6)	0.1874	0.8510	(80, 56, -2, 6)	0.1868	0.8502	72.44	(90, 50, 2, 5)	0.1985	0.8502	65.19
	0.6	(48, 5)	0.1742	0.8558	(47, 32, 2, 4)	0.1968	0.8504	37.27	(49, 25, 1, 4)	0.1995	0.8516	35.62
0.45	0.6	(80, 6)	0.1911	0.8510	(80, 56, -2, 6)	0.1905	0.8502	72.38	(94, 49, 2, 5)	0.1993	0.8509	66.12
	0.65	(47, 5)	0.1754	0.8501	(47, 32, 2, 4)	0.1995	0.8520	37.30	(53, 23, 1, 4)	0.1988	0.8519	36.24
0.5	0.65	(80, 6)	0.1923	0.8535	(80, 53, -1, 6)	0.1898	0.8503	69.59	(94, 48, 2, 5)	0.1984	0.8507	65.47
	0.7	(47, 5)	0.1767	0.8539	(47, 31, 2, 4)	0.1973	0.8515	36.63	(50, 23, 1, 4)	0.1984	0.8503	34.92
0.55	0.7	(78, 6)	0.1881	0.8503	(78, 58, -2, 6)	0.1878	0.8501	71.59	(85, 49, 2, 5)	0.1994	0.8508	62.70
	0.75	(46, 5)	0.1728	0.8533	(44, 25, 1, 4)	0.1995	0.8512	33.43	(45, 24, 1, 4)	0.1976	0.8500	33.29
0.6	0.75	(77, 6)	0.1828	0.8542	(75, 51, 2, 5)	0.1998	0.8514	60.14	(77, 41, 1, 5)	0.1997	0.8501	57.39
	0.8	(45, 5)	0.1664	0.8557	(39, 28, 1, 4)	0.1991	0.8511	32.90	(41, 18, 0, 4)	0.1959	0.8500	31.05
0.65	0.8	(74, 6)	0.1715	0.8523	(66, 36, -1, 5)	0.1999	0.8507	55.34	(71, 32, 0, 5)	0.1941	0.8510	53.53
	0.85	(36, 4)	0.1935	0.8502	(36, 27, -1, 4)	0.1932	0.8501	32.99	(39, 16, 0, 4)	0.1844	0.8504	29.19
0.7	0.85	(62, 5)	0.1888	0.8534	(62, 36, -1, 5)	0.1854	0.8502	52.91	(69, 38, 2, 4)	0.1975	0.8503	48.96
	0.9	(34, 4)	0.1769	0.8503	(34, 23, -1, 4)	0.1763	0.8500	30.54	(42, 14, 1, 3)	0.1978	0.8500	25.71
0.75	0.9	(58, 5)	0.1670	0.8532	(50, 29, 0, 4)	0.1994	0.8505	40.77	(54, 20, 0, 4)	0.1933	0.8505	39.47
	0.95	(32, 4)	0.1555	0.8568	(26, 11, 0, 3)	0.1949	0.8518	19.97	(27, 8, 0, 3)	0.1906	0.8516	19.68
0.8	0.95	(45, 4)	0.1774	0.8582	(42, 29, 2, 3)	0.1976	0.8506	33.03	(46, 19, 1, 3)	0.1974	0.8510	30.31
0.85	0.95	(85, 5)	0.1662	0.8515	(73, 42, 0, 4)	0.1979	0.8501	59.40	(79, 27, 0, 4)	0.1903	0.8500	56.97

to the single stage, the minimax design saves about 11 patients per arm in the expected sample size under H_0 .

In fact, the CALGB study chose the optimal design under a slightly larger type I error $\alpha^* = 0.16$: $(n_1, n, a_1, a) = (27, 63, 1, 5)$, which has $(\alpha, 1 - \beta, \text{EN}) = (0.1593, 0.8006, 42.87)$. With an increase of 1 per cent in α^* , we drastically reduce n and EN.

3. EXTENSIONS

So far, we have considered two-arm randomized phase II trials by allocating equal number of patients to each arm. Also, we have controlled type I and II errors under point null and alternative hypotheses. In this section, we investigate some extensions from these standard design settings.

3.1. Unbalanced randomized trials

One may want to accrue more patients to one arm than the other for some reasons, e.g. to collect more information on one arm than the other or to collect enough specimens for a correlative study on one arm. Suppose that we wish to randomize a different number of patients between two arms. Let m_l and n_l denote the sample sizes at stage l ($l = 1, 2$) of Arms 1 and 2, respectively ($m = m_1 + m_2, n = n_1 + n_2$). Also, let X_l and Y_l denote the number of responders among stage l patients of Arms 1 and 2, respectively ($X = X_1 + X_2, Y = Y_1 + Y_2$). If we want to assign r times larger number of patients to Arm 1 than to Arm 2, then we have $m_l = r \times n_l$ and $m = r \times n$. Note that a choice of $r = 1$ corresponds to the balanced two-stage designs considered in the previous section. When $r \neq 1$, it does not make sense to directly compare the numbers of responders between arms at each stage. Instead, we propose to compare the sample response rates between arms.

A two-stage design under an unbalanced allocation scheme proceeds as follows. At the first stage, we accrue m_1 patients to Arm 1 and n_1 patients to Arm 2. We continue to the second stage, if we have $X_1/m_1 - Y_1/n_1 \geq a_1$ for a constant $a_1 \in [-1, 1]$. Otherwise, we reject Arm 1 (or fail to reject H_0) and stop the trial. At the second stage, we accrue an additional m_2 patients to Arm 1 and n_2 patients to Arm 2. If, for a constant $a \in [-1, 1]$, we have $X/m - Y/n \geq a$, then we accept Arm 1 for further investigation (or reject H_0). Otherwise, we reject Arm 1. Given $H_0: p_1 = p_2 = p_0$ and $H_1: (p_1, p_2) = (p_0 + \Delta, p_0)$, the type I error and power of the two-stage design are calculated as

$$\alpha = P(X_1/m_1 - Y_1/n_1 \geq a_1, X/m - Y/n \geq a | p_1 = p_2 = p_0)$$

and

$$1 - \beta = P(X_1/m_1 - Y_1/n_1 \geq a_1, X/m - Y/n \geq a | p_1 = p_0 + \Delta, p_2 = p_0)$$

respectively.

When H_0 is true, the probability of early termination and the expected sample size for Arm 1 are calculated as

$$\text{PET} = P(X_1/m_1 - Y_1/n_1 < a_1 | p_0) = \sum_{x_1=0}^{m_1} \sum_{y_1=0}^{n_1} I(x_1/m_1 - y_1/n_1 < a_1) b(x_1 | m_1, p_0) b(y_1 | n_1, p_0)$$

and

$$EN_1 = m_1 \times PET + m(1 - PET)$$

respectively. Note that the expected sample size for Arm 2 under H_0 is obtained as $EN_2 = EN_1/r$, and the total expected sample size is $EN = EN_1 + EN_2 = (r + 1)EN_2$. Among the two-stage designs satisfying the (α^*, β^*) -restriction, the 'optimal design' is defined as the one with the smallest EN_1 (or EN).

The 'minimax design' is defined as the one with the smallest m (or $m + n$) among the two-stage designs satisfying the (α^*, β^*) -restriction.

Example 2

In CALGB 50502, suppose that we wish to assign twice as many patients to SGN-30 plus GVD arm (Arm 1), i.e. $r = 2$. Under the same design setting as in Example 1, $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.7, 0.85, 0.15, 0.8)$, the minimax design is $(m_1, m, n_1, n, a_1, a) = (72, 96, 36, 48, 0.0556, 0.0833)$, which has $(\alpha, 1 - \beta, EN) = (0.1498, 0.8021, 118.78)$, and the optimal design is $(m_1, m, n_1, n, a_1, a) = (40, 106, 20, 53, 0.0250, 0.0755)$, which has $(\alpha, 1 - \beta, EN) = (0.1478, 0.8007, 96.06)$. Unbalanced designs usually require larger sample sizes than balanced designs. For example, for the minimax designs, the total maximal sample size for this unbalanced design, $m + n = 144$, is larger than that for the balanced design, 126 from Example 1.

Example 3

CALGB 50401 randomizes non-HL patients who relapsed from a rituximab-containing combination regimen to rituximab + lenalidomide (Arm 1) and rituximab alone (Arm 2) with 1-to-2 allocation ($r = \frac{1}{2}$). Arm 2 regimen is a potential control arm for a future phase III trial in case Arm 1 is accepted in this trial, but there are not enough historical data on Arm 2. So, Arm 2 accrues twice more patients for better estimation of the clinical parameters to be used in designing a future phase III trial. The investigators would not be interested in the combination regimen (Arm 1) if its true response rate is 15 per cent or lower. Suppose that we want to test $H_0 : p_1 = p_2 = 0.15 (= p_0)$ versus $H_1 : (p_1, p_2) = (0.3, 0.15)$. Under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.15, 0.3, 0.15, 0.8)$, the minimax design is $(m_1, m, n_1, n, a_1, a) = (24, 34, 48, 69, -0.0208, 0.0759)$, which has $(\alpha, 1 - \beta, EN) = (0.1463, 0.8001, 91.78)$, and the optimal design is $(m_1, m, n_1, n, a_1, a) = (23, 35, 47, 71, 0.0037, 0.0736)$, which has $(\alpha, 1 - \beta, EN) = (0.1456, 0.8003, 87.37)$. Since our search program for minimax and optimal designs goes through all possible combinations of $n = 2m$ and $n = 2m \pm 1$, we actually have $m \approx r \times n$. In fact, CALGB 50401 was designed before the topic of this paper was developed, so this study was designed for independent evaluation of each arm as in a single-arm trial with a historical control.

3.2. Strict type I and II error control

So far, we have considered a point null hypothesis $H_0 : p_1 = p_2 = p_0$ based on the response rate for a historical control, p_0 . However, possibly due to a slightly different patient population or the variability of the estimated response rate for a historical control, the true response rate for the prospective control of a randomized trial may be different from p_0 . In this case, the chosen critical values (a_1, a) under the point null hypothesis may not control the type I error accurately under the composite null hypothesis $H_0 : p_1 = p_2$. In order to protect the type I error probability accurately

under the composite null hypothesis, we propose to calculate type I error as

$$\alpha = \max_{p_0 \in [0,1]} P(X_1 - Y_1 \geq a_1, X - Y \geq a | p_1 = p_2 = p_0) \quad (1)$$

Because $B(n, p)$ has the largest variance with $p = \frac{1}{2}$, the probability in (1) is maximized at $p_0 = \frac{1}{2}$. Hence, (1) is simplified to

$$\alpha = P(X_1 - Y_1 \geq a_1, X - Y \geq a | p_1 = p_2 = \frac{1}{2})$$

We have also considered a point alternative hypothesis for power calculation. So, a chosen two-stage design based on the point alternative hypothesis may be underpowered although the experimental arm really has a response rate higher than the control by Δ , i.e. $H_1 : p_1 = p_2 + \Delta$. In order to guarantee a certain power level over the composite alternative hypothesis, we may calculate the power by

$$1 - \beta = \min_{p_0 \in [0, 1-\Delta]} P(X_1 - Y_1 \geq a_1, X - Y \geq a | p_1 = p_0 + \Delta, p_2 = p_0)$$

which can be simplified to

$$1 - \beta = P(X_1 - Y_1 \geq a_1, X - Y \geq a | p_1 = \frac{1}{2} + \Delta/2, p_2 = \frac{1}{2} - \Delta/2)$$

In summary, given Δ , if a design (n_1, n, a_1, a) has type I error α under $H_0 : p_1 = p_2 = \frac{1}{2}$ and power $1 - \beta$ under $H_1 : p_1 = \frac{1}{2} + \Delta/2, p_2 = \frac{1}{2} - \Delta/2$, its type I error and power are given as α and $1 - \beta$ under the composite hypotheses $H_0 : p_1 = p_2$ and $H_1 : p_1 = p_2 + \Delta$.

Given $(\Delta, \alpha^*, \beta^*)$, the optimal and minimax designs are defined as in Section 3. We do not specify p_0 in designing a study controlling the type I error and power under composite hypotheses. For example, for $(\Delta, \alpha^*, 1 - \beta^*) = (0.15, 0.15, 0.8)$ as in Example 1, the minimax design is $(n_1, n, a_1, a) = (54, 78, -2, 7)$, which has $(\alpha, 1 - \beta, EN) = (0.1487, 0.8000, 70.43)$, and the optimal design is $(n_1, n, a_1, a) = (39, 89, 1, 7)$, which has $(\alpha, 1 - \beta, EN) = (0.1428, 0.8001, 61.75)$. Note that these sample sizes are larger than those in Example 1, which are calculated under point null and alternative hypotheses.

3.3. Randomized trials with one control and K experimental arms

Suppose that there are $K (\geq 2)$ experimental arms and one control. We wish to identify the experimental arms whose response rate is significantly higher than that of the control arm. We consider balanced allocations here, but the following results can be easily modified for an unbalanced allocation case as in Section 3.1.

In the first stage, we accrue n_1 patients to each of $K + 1$ arms. For stage 1, let X_{k1} denote the number of responders from the experimental arm $k (= 1, \dots, K)$ and Y_1 the number of responders from the control arm. For an integer $a_1 \in [-n_1, n_1]$, experimental arm k with $X_{k1} - Y_1 \geq a_1$ proceeds to the second stage together with the control. All experimental arms with $X_{k1} - Y_1 < a_1$ will be dropped because of lack of efficacy. If no experimental arm survives stage 1, then the whole trial will be terminated after stage 1. If any experimental arms survive stage 1, then the control arm will be included in stage 2.

In the second stage, we accrue an additional n_2 patients to each of the experimental arms that survived the first stage and the control. Let X_{k2} and Y_2 denote the number of responders from the second stage patients of experimental arm k and the control, respectively. Note that the number of experimental arms in the second stage may be smaller than K . Also, let $X_k = X_{k1} + X_{k2}$ and $Y = Y_1 + Y_2$ denote the total number of responders from the cumulative $n = n_1 + n_2$ patients for experimental arm k and the control, respectively. For an integer $a \in [a_1 - n_2, n]$, we accept experimental arm k for further investigation if $X_k - Y \geq a$.

Let p_k denote the response rate for the experimental arm $k (= 1, \dots, K)$, and q that for the control arm. Also, let p_0 denote the response rate for a historical control. We consider point null hypothesis $H_0 : p_1 = \dots = p_K = q = p_0$. We propose to control the probability of erroneously accepting any inefficacious experimental arm, called family-wise error rate (FWER),

$$\begin{aligned} \alpha &= P \left\{ \bigcup_{k=1}^K (X_{k1} - Y_1 \geq a_1, X_k - Y \geq a) \mid p_0 \right\} \\ &= \sum_{y_1=0}^{n_1} \sum_{x_{11}=0}^{n_1} \dots \sum_{x_{K1}=0}^{n_1} \sum_{y_2=0}^{n_2} \sum_{x_{12}=0}^{n_2} \dots \sum_{x_{K2}=0}^{n_2} I \left\{ \bigcup_{k=1}^K (x_{k1} - y_1 \geq a_1, x_{k1} + x_{k2} - y_1 - y_2 \geq a) \right\} \\ &\quad \times b(y_1 | n_1, p_0) b(y_2 | n_2, p_0) \prod_{k=1}^K b(x_{k1} | n_1, p_0) b(x_{k2} | n_2, p_0) \end{aligned} \quad (2)$$

The family-wise power under a specified alternative hypothesis $H_1 : q = p_0, p_k = p_0 + \Delta (k = 1, \dots, K)$ is calculated as

$$\begin{aligned} 1 - \beta &= P \left\{ \bigcup_{k=1}^K (X_{k1} - Y_1 \geq a_1, X_k - Y \geq a) \mid q = p_0, p_1 = \dots = p_K = p_0 + \Delta \right\} \\ &= \sum_{y_1=0}^{n_1} \sum_{x_{11}=0}^{n_1} \dots \sum_{x_{K1}=0}^{n_1} \sum_{y_2=0}^{n_2} \sum_{x_{12}=0}^{n_2} \dots \sum_{x_{K2}=0}^{n_2} I \left\{ \bigcup_{k=1}^K (x_{k1} - y_1 \geq a_1, x_{k1} + x_{k2} - y_1 - y_2 \geq a) \right\} \\ &\quad \times b(y_1 | n_1, p_0) b(y_2 | n_2, p_0) \prod_{k=1}^K b(x_{k1} | n_1, p_0 + \Delta) b(x_{k2} | n_2, p_0 + \Delta) \end{aligned} \quad (3)$$

Given $(p_0, \Delta, \alpha^*, \beta^*)$, the optimal and minimax designs are defined as in a two-arm trial case. Let us consider the case where $K = 2$. There are two types of early termination: (i) when only one experimental arm is rejected, or (ii) when both experimental arms are rejected after stage 1. For type (i), the required sample size is $3n_1 + 2n_2$ and the probability of early termination under H_0 is

$$\text{PET}_1 = 2 \times P(X_{11} - Y_1 < a_1, X_{21} - Y_1 \geq a_1 | p_0)$$

and, for type (ii), the required sample size is $3n_1$ and the probability of early termination under H_0 is

$$\text{PET}_2 = P(X_{11} - Y_1 < a_1, X_{21} - Y_1 < a_1 | p_0)$$

Hence, the expected sample size under H_0 is obtained as

$$\begin{aligned} \text{EN} &= (3n_1 + 2n_2)\text{PET}_1 + 3n_1 \times \text{PET}_2 + 3n(1 - \text{PET}_1 - \text{PET}_2) \\ &= 3n - n_2 \times \text{PET}_1 - 3n_2 \times \text{PET}_2 \end{aligned}$$

$\frac{1}{3}$ of which is the expected sample size per arm.

Even with $K = 2$, the search for the optimal and minimax designs requires heavy computations. For an expedited search, we may choose a reasonable n , e.g. an integer slightly larger than that for a two-arm design, and find (n_1, a_1, a) that satisfy (α^*, β^*) in a narrow space, such as $n_1 \in [0.3n, 0.7n]$, $a_1 \in [-2, 2]$ and $a \in [n\Delta/2 - 2, n\Delta/2 + 2]$. This suggestion is based on our experience that an n_1 around $n/2$ provides a convenient time schedule for the interim analysis, and, for reasonable two-stage designs, a_1 is chosen around 0 and a is chosen around $n\Delta/2$.

Example 4

Let us consider $(p_0, \Delta, \alpha^*, 1 - \beta^*) = (0.7, 0.15, 0.15, 0.8)$ and $K = 2$. We may choose $n = 70$, which is slightly larger than that for the minimax design for two-arm trials, 63 from Example 1, $n_1 \in [21, 49]$, $a_1 \in [-2, 2]$ and $a \in [3, 8]$. Within the range, we choose the design with the smallest EN among those satisfying the $(\alpha^*, 1 - \beta^*)$ -restriction. From the expedited search, we find design $(n_1, n, a_1, a) = (23, 70, 2, 7)$, which has operating characteristics $(\alpha, 1 - \beta) = (0.1382, 0.8003)$ and $\text{EN} = 40.01$ per arm.

In order to adjust for the multiplicity of statistical tests, we propose to control the FWER in testing and to choose a design satisfying the family-wise power $1 - \beta$ given in (3). However, one may want to choose a design satisfying the marginal power to accept each efficacious experimental therapy with a certain probability. Given (n_1, n) , suppose that the critical values (a_1, a) are chosen to control the FWER given in (2) below α^* level. Then, the marginal power for experimental arm k with $p_k = p_0 + \Delta$ will be calculated as

$$\begin{aligned} 1 - \tilde{\beta} &= P(X_{k1} - Y_1 \geq a_1, X_k - Y \geq a | p_0, p_k) \\ &= \sum_{y_1=0}^{n_1} \sum_{x_{k1}=0}^{n_1} \sum_{y_2=0}^{n_2} \sum_{x_{k2}=0}^{n_2} I(x_{k1} - y_1 \geq a_1, x_{k1} + x_{k2} - y_1 - y_2 \geq a) \\ &\quad \times b(y_1 | n_1, p_0) b(y_2 | n_2, p_0) b(x_{k1} | n_1, p_k) b(x_{k2} | n_2, p_k) \end{aligned}$$

In Example 4 with $K = 2$, the design $(n_1, n, a_1, a) = (23, 70, 2, 7)$ has a marginal power of $1 - \tilde{\beta} = 0.6654$ for $(p_0, \Delta) = (0.7, 0.15)$. If we want $(\alpha^*, 1 - \tilde{\beta}^*) = (0.15, 0.8)$ for each experimental arm with $(p_0, \Delta) = (0.7, 0.15)$, then we need a larger trial, such as $(n_1, n, a_1, a) = (44, 88, 0, 9)$ which has $(\alpha, 1 - \tilde{\beta}) = (0.1293, 0.8007)$.

4. DISCUSSION

The patient characteristics of the experimental arm of a new single-arm phase II trial may be different from those of the trial from which the historical control is selected. In this case, the

comparison between the experimental arm and the historical control may be biased. We avoid such issue by randomizing patients between the experimental and control therapies.

While the number of randomized phase II trials is rapidly growing [11], we largely lack efficient design and analysis methods for them. This paper proposes optimal and minimax designs for two-stage randomized phase II trials. Given a design setting, the maximal sample size for the minimax two-stage design is usually smaller than or equal to the sample size for the single-stage design as in single-arm trial designs [1]. The ratio of stage 1 sample size to the maximal size, n_1/n , for the minimax design is usually large, so that its operating characteristics and the maximal sample size are similar to those of the single-stage design. However, the ratio for the optimal design is usually small, so that we can terminate the trial early and minimize the expected sample size when the experimental arm is inefficacious.

We have considered minimax and optimality criteria in this paper. But these two criteria often conflict with each other, so that the minimax design may have an excessively large expected sample size under H_0 compared with the optimal design and the optimal design may have an excessively large maximal sample size compared with the minimax design. In order to address this issue, we may combine these two criteria to derive a compromise design; refer to Jung *et al.* [3, 12] for the single-arm design case. We have focused on two-stage designs, but the methods can be easily extended to designs with any number of stages.

A randomized phase II trial may look similar to a phase III trial in the sense that both include a prospective control and carry out statistical tests to compare between the control and an experimental arm. However, we do not want a phase II trial to be more than an efficacy screening study, while a phase III trial is to finalize scientific questions on an experimental regimen. As a result, we want phase II trials as simple as possible. In order to keep the sample size small and the study period short for a randomized phase II trial, we use relatively large α , such as 15 or 20 per cent (rather than the conventional 5 per cent level), and β , such as 20 per cent (rather than 10 per cent), and a short-term outcome variable, such as tumor response (rather than survival), as the primary endpoint.

When there are multiple strata, among which the response rates are different, a referee wishes to consider a separate single-arm phase II trial for each stratum as an alternative to a randomized phase II trial. Through such types of studies, we may avoid the bias issue, but may face some different practical problems. Usually in this situation, medical investigators might want to decide whether a new therapy has promising efficacy or not for the whole group of patients combining the strata. Suppose that, by a study with separate single-arm phase II trials, the experimental therapy is accepted in one stratum, but not in another by a marginal difference in efficacy. In this case, we may want to accept the experimental therapy for the combined population. But it is not clear what the type I error probability committed by the decision is. If the strata are so different that they are likely to have different standard therapies, then we will have to consider a separate single-arm phase II trial for each stratum as the referee suggests. Compared with randomized phase II trials, studies with multiple single-arm phase II trials may not save the sample size at all, especially when there are more than two strata. Furthermore, the whole study will not be completed until all strata accrue the required number of patients, so that it may take a long time to complete a study if there is a rare stratum.

Finally, with respect to computing, the computer programs used in this paper are written in Fortran 77. It takes about a minute to search for the minimax and optimal designs under each setting of Tables I–IV by a laptop computer with Intel Pentium M Processor 1.8 GHz. The computing time exponentially increases in K , so that an exhaustive search can take many days even for a

three-arm phase II trial ($K = 2$). Using the expedited search procedure discussed in Section 3.3, we can identify reasonable designs for a three-arm trial within 10 min. The computer programs are available from the author on request.

REFERENCES

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
2. Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine* 2004; **23**:881–896.
3. Jung SH, Lee TY, Kim KM, George S. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 2004; **23**:561–569.
4. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treatment Reports* 1985; **69**:1375–1381.
5. Liu PY, LeBlanc M, Desai M. False positive rates of randomized phase II designs. *Controlled Clinical Trials* 1999; **20**:343–352.
6. Sargent DJ, Goldberg RM. A flexible design for multiple armed screening trials. *Statistics in Medicine* 2001; **20**:1051–1060.
7. Thall PF, Simon R, Ellenberg SS. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* 1989; **45**:537–547.
8. Palmer CR. A comparative phase II clinical trials procedure for choosing the best of three treatments. *Statistics in Medicine* 1991; **10**:1327–1340.
9. Steinberg SM, Venzon DJ. Early selection in a randomized phase II clinical trial. *Statistics in Medicine* 2002; **21**:1711–1726.
10. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. *Journal of Clinical Oncology* 2005; **23**(28):7199–7206.
11. Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. *Journal of Clinical Oncology* 2005; **23**(19):4450–4457.
12. Jung SH, Carey M, Kim KM. Graphical search for two-stage designs for phase II clinical trials. *Control Clinical Trials* 2001; **22**:367–372.