Journal of Biometrics & Biostatistics
ISSN: 2155-6180

Chang et al., J Biom Biostat 2015, 6:2
DOI: 10.4172/2155-6180.1000229

Research Article Open Access

Alternative Designs for Phase II Clinical Trials when Attained Sample Sizes are Different from Planned Sample Sizes

Chang M*, Li Y and An Q

Department of Biostatistics, University of Florida, USA

Abstract

Phase II cancer clinical trials are conducted for initial evaluation of the therapeutic efficacy of a new treatment regimen, and two-stage designs often are implemented in such trials. Typically, designs of phase II trials not only satisfy predefined significance and power requirements but also have some desirable features such as minimizing the total sample size, minimizing the average sample size under the null hypothesis, etc. A frequent issue is that the attained sample sizes differ from the planned sample sizes. We propose alternative designs adjusted to the attained sample sizes when they are different than the planned sample sizes. We present extensive examples and compare the proposed designs to that of Green and Dahlberg. We apply the proposed designs to a phase II trial in non-Hodgkin's lymphoma patients.

Keywords: Alternative designs; Phase II trial; Sample sizes; Clinical trails

Introduction

In a phase II cancer clinical trial, investigators seek to determine whether a new treatment has sufficient promise for further study in a large-scale phase III trial. After the determination of appropriate dose level in Phase I trials, phase II clinical trials are undertaken to provide an initial assessment of the treatment efficacy, typically in terms of response rate of the treatment. Usually, the response rate of the standard treatment is known, and a treatment is considered promising if its response rate is better than the standard level by some predefined margin. This can be set up as testing the null hypothesis

$$H_0: p \le p_0 \text{ vs. the alternative hypothesis } H_1: p > p_1, \quad (1)$$

Where p is the response rate of the experimental treatment, p_0 is the response rate of the standard treatment, and p_1 is the level of response at which one considers the treatment promising. For ethical and efficiency reasons, most phase II trials use sequential designs. Two-stage designs are commonly used because of their logical simplicity and the diminishing benefit of multistage trials over two stages. Many designs for phase II clinical trials have been proposed [1-7]. The proposed designs usually not only satisfy predefined significance and power requirements but also have some desirable features such as minimizing the total maximum number of patients, or minimizing the average number of patients under the null hypothesis.

A frequent issue is that the accrual in stages 1 and 2 may not proceed exactly as planned. As pointed by Green and Dahlberg [8], there are logistic problems encountered in the conduct of multicenter studies, including the following:

1. Accrual cannot be suspended immediately after enrolling a specified number of patients, as institutions may be allowed to enroll patients for whom recruitment efforts have already begun.

2. Communication of information such as study closure is often slow in large bureaucracies or multicenter study groups such as a cancer cooperative clinical trial group.

In addition, investigators may find that some patients who entered the study are ineligible after the accrual is suspended at either stage 1 or stage 2. Consequently, the number of evaluable patients is smaller than originally expected. As pointed by Herndon [9], the accrual suspension is unfeasible in some studies because of long treatment durations and/or long time periods required for response assessment. If the decision of whether the study should be stopped is made during a meeting of the Data Safety Board, the attained sample size will likely differ from the planned sample size. The question is how the design should be modified when the attained sample sizes are different than the planned sample sizes.

For example, suppose we want to evaluate the efficacy of lenalidomide in a phase II trial in non-Hodgkin's lymphoma patients who responded and then relapsed after the first chemotherapy. A response rate of p_0=0.20 or lower is considered too low and a response rate of p_1=0.40 or higher is considered promising. Assume that investigators plan to accrue 20 patients at the first stage. If the number of responses among the 20 patients is 4 or less, then the study is terminated and the treatment is rejected. Otherwise, an additional 20 patients enter the study at the second stage. If the number of responses among the total of 40 patients is 11 or less by the end of stage 2, then the treatment is rejected. Otherwise, the treatment is considered promising. Assume that the attained sample sizes are 18 and 38 in stages 1 and 2, respectively. What alternative design should the investigators use? We will address this problem below.

Two methods were proposed by Green and Dahlberg [8] and Hernden [9]. Green and Dahlberg [8] proposed to conduct a one-sided test of the alternative hypothesis H_1: $p>p_1$ at the 0.02 level in stage 1 and conduct a one-sided test at the 0.055 significance level in stage 2. Their simple and uniform approach is attractive, and their designs have been used until recently [10]. Hernden [9] proposed hybrid designs for some phase II clinical trials without accrual suspension during the evaluation of response status of patients entered for hypothesis testing at stage 1. In general, the method of error rate spending function can

***Corresponding author:** Chang M, Department of Biostatistics, University of Florida, USA, Tel: 352-294-5914; FAX: 352-294-5931; E-mail: mchang@biostat.ufl.edu

Received April 04, 2015; **Accepted** June 02, 2015; **Published** June 09, 2015

Citation: Chang M, Li Y, An Q (2015) Alternative Designs for Phase II Clinical Trials when Attained Sample Sizes are Different from Planned Sample Sizes. J Biom Biostat 6: 229. doi:10.4172/2155-6180.1000229

be applied for group sequential testing with unpredictable sample sizes. Lan and DeMets [11] proposed a flexible sequential testing procedure by selecting a type I error probability spending function before the study starts. Pampallona et al. in a Harvard School of Public Health technical report and Chang et al. [12] extended the method in group sequential testing using both type I and type II error probability spending functions. Further development can be found in Hampson and Jennison [13].

In our first approach, we define type I and type II error probability spending functions by the planned design, and then conduct two-stage testing using these error probability spending functions according to the attained sample sizes. In designs of Green and Dahlberg [8], the type II error probability spent at the first stage is uniformly set up at a fixed level of 0.02. Their designs may be quite different with the panned designs and may not have the desirable features as in the panned designs. Our proposed designs are more flexible, closer to the planned design, and preserve the desirable features as in the planned design better than designs of Green and Dahlberg [8]. Our second approach is to redesign the two-stage testing procedure after the sample size at the first stage is attained, following the same criteria as in the planned designs. If the attained sample size is different than the redesigned sample size at the second stage, then the threshold at the second stage will be adjusted to satisfy the significance requirement. Our designs satisfy the requirement on significance level. In addition, the power of the proposed design is close to the nominal level when the total sample size is close to that in the planned sample size. Other parameters in the proposed design, such as average sample size, are close to those in the planned design as well.

In next section, we summarize typical planned designs for two-stage phase II clinical trials. In the section following the next section, we introduce alternative designs when attained sample sizes are different than planned sample sizes. Then we present numerical examples and a real example. The last section is a brief discussion.

Planned Designs

The typical objective of phase II clinical trials is to evaluate experimental treatments that might increase the response rate over a historical level. We set up the null and alternative hypotheses as in (1). A phase II clinical trial is usually carried out in two stages. We classify most popular designs in two-stage clinical trials as category I and category II as follows. A category I design allows early termination of the trial at the end of the first stage when the treatment is ineffective, and allows the trial to continue to the second stage when the data indicate that the treatment has a certain efficacy. A category II design allows early termination of the trial not only when the treatment is ineffective but also when the treatment is clearly effective at the first stage. We denote sample sizes at stages 1 and 2 and the total sample size by n_1, n_2, and $n=n_1+n_2$, respectively, and denote the numbers of responses at stages 1 and 2, and the cumulative number of responses across two stages, by Y_1, Y_2, and $Y=Y_1+Y_2$, respectively.

A category I design is specified by a threshold a at the first stage and a threshold c at the second stage. After n_1 patients enter the study and their response statuses are evaluated, we conduct the test at the first stage as follows: the study is terminated if $Y_1 \le a$, and the treatment is rejected; if $Y_1>a$, then the testing procedure is continued to the second stage. At the second stage, the treatment is rejected or considered promising when $Y \le c$ or $Y>c$, respectively. The significance level of the test is

$$P(Y_1 > a, Y > c \mid n_1, n_2, p = p_0); \quad (2)$$

the power is

$$P(Y_1 > a, Y > c \mid n_1, n_2, p = p_1); \quad (3)$$

and the average sample size under the null hypothesis is

$$Aven = n_1 + n_2 P(Y_1 > a \mid n_1, p = p_0) \quad (4)$$

We denote a category I design by its parameters (n_1, n_2, a, c). Simon's minimax designs [6] are category I designs that minimize the total sample size n under the requirements of significance level and power: (2) $\le \alpha$ and (3)$>1-\beta$. Simon's optimal designs [6] are category I designs that minimize the average sample size in (4) under the constraints (2) $\le \alpha$ and (3)$>1-\beta$.

A category II design is specified by adding a threshold b at the first stage in addition to design parameters (n_1, n_2, a, c) as in a category I design. After n_1 patients enter the study and their response statuses are evaluated, we conduct the test at the first stage as follows: the study is terminated if $Y_1 \le a$ or $Y_1>b$, and the treatment is rejected or considered promising correspondingly; if $a<Y_1<b$, then the testing procedure is continued to the second stage. At the second stage, the treatment is rejected or considered promising when $Y \le c$ or $Y>c$, respectively. The significance level of the test is

$$P(Y_1 \ge b \mid n_1, p = p_0) + P(a < Y_1 < b, Y > c \mid n_1, n_2, p = p_0); \quad (5)$$

the power is

$$P(Y_1 \ge b \mid n_1, p = p_1) + P(a < Y_1 < b, Y > c \mid n_1, n_2, p = p_1); \quad (6)$$

and the average sample size under the null hypothesis is

$$Aven = n_1 + n_2 P(a < Y_1 < b \mid n_1, p = p_0) \quad (7)$$

We denote a category II design by its parameters (n_1,n_2,a,c).The minimax design is the design that minimizes the total sample size n under the constraints (5)$<\alpha$ and (6)$>1-\beta$. The optimal design is the design that minimizes the average sample size in (7) under the constraints (5) $\le \alpha$ and (6)$>1-\beta$.

The minimax or the optimal design can be obtained by a search program. Compared with a category II design, a category I design has a higher probability of letting the trial proceed to the second stage, and provides a more accurate estimation of the response rate or other endpoints when the treatment is effective. In contrast, a category II design allows early termination and shortens the duration of the trial when the treatment is either clearly effective or ineffective. If we define the rejection region of the null hypothesis at stage 1 to be a null set in a category II design ($b=n_1+1$), then the category II design becomes a category I design. Therefore, category I designs are special cases of category II designs. A minimax or an optimal design of category II is more efficient than that of a category I design since category I designs form a subset of category II designs.

A common approach is to use equal or almost equal sample sizes between the two stages ($n_1=n_2$ or $n_1=n_2+1$). The design with equal sample sizes or almost equal sample sizes has the advantage of simplicity: formal data monitoring is performed after half or about half of the patients enter the study. The minimax and the optimal design can be obtained by a search program under the constraint: $n_1=n_2$ or $n_1=n_2+1$. Some clinicians do not pursue the optimal property in category I or category II designs. They intend to have the testing procedure proceed to the second stage unless the result at the first stage is extreme. They select a low value of a in a category I design or select a low value of

a and a large value of *b* in a category II design, such that the probability of early stopping is low. In this way, the data will likely provide a more accurate estimate of the response rate and other endpoints.

Alternative Designs

A frequent issue is that the attained sample sizes at stages 1 and 2, denoted by n_1^* and n_2^* ($n^* = n_1^* + n_2^*$), are different than the planned sample sizes n_1 and n_2. We propose in Alternative designs using type I and category II error probability spending functions to conduct the testing procedure using type I error and type II error spending functions defined by the planned design. In Alternative designs with redesigns adjusted to the attained sample sizes, we propose to redesign the testing procedure after the sample size n_1^* is attained at the first stage.

Alternative designs using type I and category II error probability spending functions

For a category I design (n_1, n_2, a, c), there is no type I error probability spent at the first stage, and all type I error probability is spent at the second stage. We define the type II error probability spending function according to the planned design. We then obtain the attained design using the error probability spending functions according to the attained sample sizes n_1^* and n_2^*. Assume the required significance level and power are α and 1-β, respectively. The type II error probability spent at stage 1 is

$$\beta_1 = P(Y_1 \le a \mid n_1, p = p_1). \tag{8}$$

We define the type II error probability spending function as a piece-wise linear function of sample size *m*:

$$\beta(m) = \begin{pmatrix} \beta_1 m / n_1 & if\ m \le n_1 \\ \beta_1 + (\beta - \beta_1)(m - n_1)/n_2 & if\ m > n_1 \end{pmatrix}. \tag{9}$$

We select integer a^* as the threshold at stage 1 such that

$$P(Y_1 \le a^* \mid n_1^*, p_1) \approx \beta(n_1^*),$$

where "$\approx$" means "is closest to." We further select the smallest integer c^* as the threshold at stage 2 such that

$$P(Y_1 > a^*, Y > c^* \mid n_1^*, n_2^*, p_0) \le \alpha \ .$$

We conduct the testing procedure using the design (n_1^*, n_2^*, a^*, c^*). The proposed design satisfies the significance requirement. If the attained sample sizes are n_1^* and n_2^* close to the planned sample size n_1 and n_2, then the attained design approximately satisfies the power requirement. Since the type II error probability spending function is used, the attained design has approximately the desired features as in the planned design.

For a category II design (n_1, n_2, a, c), the type I error probability spent at stage 1 is

$$\alpha_1 = P(Y_1 \ge b \mid n_1, p = p_0) \ ,$$

and the type II error probability spent at stage 1 is in (8). We define the type I error probability spending function as a piece-wise linear function of sample size *m*:

$$\alpha(m) = \begin{pmatrix} \alpha_1 m / n_1 & if\ m \le n_1 \\ \alpha_1 + (\alpha - \alpha_1)(m - n_1)/n_2 & if\ m > n_1 \end{pmatrix} \tag{10}$$

We define the type II error probability spending function as in (9). We select integers a* and b* as the thresholds at stage 1 such that

$$P(Y_1 \ge b^* \mid n_1^*, p_0) \approx \alpha(n_1^*) \text{ and}$$

$$P(Y_1 \le a^* \mid n_1^*, p_1) \approx \beta(n_1^*).$$

We further select the smallest integer c^* as the threshold at stage 2 such that

$$P(Y_1 \ge b^* \mid n_1^*, p_0) + P(a^* < Y_1 < b^*, Y > c^* \mid n_1^*, n_2^*, p_0) \le \alpha$$

We conduct the testing procedure using the design $(n_1^*, n_2^*, a^*, b^*, c^*)$. The proposed design satisfies the significance requirement. If the attained sample sizes are n_1^* and n_2^* close to the planned sample sizes n_1 and n_2, then the attained design approximately satisfies the power requirement. Since type I and type II error probability spending functions are used, the attained design has approximately the desired features as in the planned design.

Alternative designs with redesigns adjusted to the attained sample sizes

We propose in this section to redesign the testing procedure when the attained sample size n_1^* is different than the planned sample size n_1 at the first stage. For a category I design, we determine parameters are a^*, n_2^*, and c^* conditional on n_1^* to satisfy the significance level and power requirements:

$$P(Y_1 > a^*, Y > c^* \mid n_1^*, n_2^*, p = p_0) \le \alpha \text{ and}$$

$$P(Y_1 > a^*, Y > c^* \mid n_1^*, n_2^*, p = p_1) \ge 1 - \beta \ .$$

In addition, the design (n_1^*, n_2^*, a^*, c^*) satisfies the same criteria as in the originally planned design, such as minimizing the total sample size, minimizing the average sample size under the null hypothesis, etc. The sample size n_2^* may be different than the originally planned sample size n_2. By the end of the first stage, the accrual target of the sample size at stage 2 is changed to n_2^* from n_2 if the testing procedure is continued to the second stage. By the end of the second stage, the attained sample size n_2^{**} may be different than the redesigned sample size n_2^*. Further adjustment of the threshold at the second stage is needed. For given $n_1^*, n_2^{**}, and\ a^*$, we select the smallest integer c^{**} satisfying

$$P(Y_1 > a^*, Y > c^{**} \mid n_1^*, n_2^{**}, p = p_0) \le \alpha \ .$$

The final design is $(n_1^*, n_2^{**}, a^*, c^{**})$.

We emphasize that the determination of design parameters n_2^*, a^*, c^* and c^{**} does not depend on response data at stages 1 and 2.

For a category II design and when the attained sample size n_1^* is different than the planned sample size n_1 at the first stage, we determine the parameters are a^*, b^*, n_2^*, and c^* conditional on n_1^* to satisfy the significance and power requirements:

$$P(Y_1 \ge b^* \mid n_1^*, p = p_0) + P(a^* < Y_1 < b^*, Y > c^* \mid n_1^*, n_2^*, p = p_0) \le \alpha \text{ and}$$

$$P(Y_1 \ge b^* \mid n_1^*, p = p_1) + P(a^* < Y_1 < b^*, Y > c^* \mid n_1^*, n_2^*, p = p_1) \ge 1 - \beta \ .$$

In addition, the design $(n_1^*, n_2^*, a^*, b^*, c^*)$ satisfies the same criteria as in the originally planned design. Similar to a category I design, the final design is $(n_1^*, n_2^{**}, a^*, b^*, c^{**})$, where n_2^{**} is the attained sample size and c^{**} is the adjusted threshold at the second stage; i.e. c^{**} is the smallest integer satisfying

$$P(Y_1 \ge b^* \mid n_1^*, p = p_0) + P(a^* < Y_1 < b^*, Y > c^{**} \mid n_1^*, n_2^{**}, p = p_0) \le \alpha.$$

We emphasize that the determination of design parameters a^*, b^*, n_2^*, c^* and c^{**} does not depend on response data at stages 1 and 2.

Examples and Comparison

Examples of proposed alternative designs versus planned designs

		Planned Design							Actual Sample Size		Attained Design									
											Proposed Design					Green and Dahlberg				
p_0	p_1	n_1	n	a	c	α	1-β	AVE N	n_1*	n*	a*	c*	α*	1-β*	AVE N	a*	c*	α*	1-β*	AVE N
0.05	0.20	19	38	1	3	0.090	0.90	23.7	17	36	0	3	0.098	0.94	28.1	0	3	0.098	0.94	28.1
									17	40	0	4	0.046	0.91	30.4	0	4	0.046	0.91	30.4
									21	36	0	4	0.032	0.87	30.9	0	4	0.032	0.87	30.9
									21	40	0	4	0.047	0.92	33.5	0	4	0.047	0.92	33.5
0.10	0.30	17	33	2	5	0.081	0.90	20.8	15	31	1	5	0.077	0.92	22.2	0	5	0.083	0.94	27.7
									15	35	0	6	0.055	0.93	30.9	0	6	0.055	0.93	30.9
									19	31	2	5	0.075	0.92	22.5	1	5	0.082	0.93	26.0
									19	35	2	6	0.049	0.91	23.7	1	6	0.054	0.93	28.3
0.20	0.40	20	40	4	11	0.078	0.90	27.4	18	38	3	11	0.060	0.88	28.0	2	11	0.062	0.89	32.6
									18	42	2	12	0.061	0.91	35.5	2	12	0.061	0.91	35.5
									22	38	4	11	0.061	0.88	29.3	4	11	0.061	0.88	29.3
									22	42	4	12	0.059	0.90	31.1	4	12	0.059	0.90	31.1
0.30	0.50	21	42	6	16	0.090	0.90	30.4	19	40	5	16	0.061	0.86	30.0	4	16	0.063	0.86	34.1
									19	44	4	17	0.080	0.91	36.9	4	17	0.080	0.91	36.9
									23	40	6	16	0.063	0.86	32.5	6	16	0.063	0.86	32.5
									23	44	6	17	0.079	0.91	34.8	6	17	0.079	0.91	34.8
0.40	0.60	25	49	11	23	0.098	0.90	31.4	23	47	10	23	0.068	0.87	29.9	8	23	0.080	0.91	37.7
									23	51	8	25	0.072	0.92	40.1	8	25	0.072	0.92	40.1
									27	47	11	23	0.077	0.91	34.7	10	23	0.081	0.92	37.8
									27	51	11	25	0.069	0.91	36.3	10	25	0.072	0.92	40.0
0.50	0.70	24	47	13	27	0.095	0.90	30.2	22	45	12	26	0.089	0.88	28.0	10	27	0.067	0.90	35.4
									22	49	10	29	0.074	0.93	37.8	10	29	0.074	0.93	37.8
									26	45	13	27	0.066	0.90	34.0	13	27	0.066	0.90	34.0
									26	49	13	29	0.073	0.92	35.7	13	29	0.073	0.92	35.7
0.60	0.80	20	39	12	27	0.083	0.91	27.9	18	37	10	26	0.071	0.89	28.7	10	26	0.071	0.89	28.7
									18	41	10	29	0.056	0.89	31.0	10	29	0.056	0.89	31.0
									22	37	13	26	0.071	0.89	28.8	13	26	0.071	0.89	28.8
									22	41	13	29	0.056	0.89	30.6	13	29	0.056	0.89	30.6
0.70	0.90	15	29	11	23	0.081	0.91	19.2	13	27	9	22	0.058	0.87	18.9	8	22	0.059	0.87	22.2
									13	31	8	25	0.062	0.92	24.8	8	25	0.062	0.92	24.8
									17	27	12	22	0.059	0.87	20.9	12	22	0.059	0.87	20.9
									17	31	12	25	0.061	0.91	22.4	12	25	0.061	0.91	22.4

Table 1a: Attained designs using type II error probability spending function versus planned category I optimal designs ($\alpha \le 0.10$, $1-\beta \ge 0.90$).

are presented in Tables 1A, 1B, 2A, and 2B. All planned designs satisfy the significance and the power requirements of $\alpha \le 0.10$ and $1-\beta > 0.90$, subject to the constraint of $n_1=n_2$ or $n_1=n_2+1$. The planned designs are optimal in the sense that the planned design has the minimum average sample size under the null hypothesis. We also listed the designs of Green and Dahlberg in the tables for comparison. The designs of Green and Dahlberg spend type I and type II error probabilities at stage 1 by a fixed amount of 0.02 (for type I designs, no type I error probability is spent at stage 1) and satisfy the overall significance requirement of $\alpha \le 0.10$.

Proposed Designs Using Type II Error Probability Spending Function Versus Planned Category I Designs

Examples of proposed designs using a type II error probability spending function versus planned category I designs are presented in Table 1A. For example, at the 9th row entry, the testing procedure is for the null hypothesis H_0: $p \le p_0=0.10$ versus the alternative hypothesis H_1: $p>p_1=0.30$. The planned design is $(n_1, n_2, a, c)=(17, 16, 2, 5)$ with a significance level of 0.081, a power of 0.90, and an average sample size under the null hypothesis of 20.8. The attained sample sizes are $n_1^*=19$ and $n_2^*=12$ ($n^*=31$) at stages 1 and 2, respectively. Using the type II error probability spending function defined in (9), we obtain the attained design $(n_1^*, n_2^*, a, c)=(19, 12, 2, 5)$. This design has a significance level of 0.075, a power of 0.92, and an average sample size under the null hypothesis of 22.5. The corresponding Green and Dahlberg design is $(n_1^*, n_2^*, a, c)=(19, 12, 1, 5)$. This design has a significance level of 0.082, a power of 0.93, and an average sample size under the null hypothesis of 26.0. In all 32 cases we investigated (Table 1a), the average performance of the proposed designs on significance and power is almost identical to that of Green and Dahlberg (mean significance level: 0.065 vs. 0.066; mean power: 0.90 vs. 0.90). In fact, in 22 cases out of the 32, the proposed designs are the same as those of Green and Dahlberg. The agreement between the proposed designs and the designs of Green and Dahlberg is due to the fact that the type II error probability computed by (9) is close to 0.02 in many cases. In the remaining 10 cases, the proposed designs uniformly have smaller average sample size under the null hypothesis than those of Green and Dahlberg (mean average sample size: 27.4 vs. 32.2).

Proposed designs using type I and II error probability spending functions versus planned category II designs

Examples of proposed designs using type I and II error probability spending functions versus planned category II designs are presented in Table 1b. For example, at the 9th row entry, the testing procedure is for

		Planned Design								Actual Sample Size		AttainedDesign											
												Proposed Design						Green and Dahlberg					
p_0	p_1	n_1	n	a	b	c	α	1-β	AVE N	n_1*	n*	a*	b*	c*	α*	1-β*	AVE N	a*	b*	c*	α*	1-β*	AVE N
0.05	0.20	19	38	1	4	3	0.090	0.90	23.4	17	36	0	4	3	0.098	0.94	27.9	0	4	3	0.098	0.94	27.9
										17	40	0	4	4	0.048	0.91	30.2	0	4	4	0.048	0.91	30.2
										21	36	0	4	4	0.039	0.88	30.6	0	4	4	0.039	0.88	30.6
										21	40	0	4	4	0.053	0.92	33.2	0	4	4	0.053	0.92	33.2
0.10	0.30	17	33	2	5	5	0.084	0.91	20.5	15	31	1	5	5	0.079	0.92	22.0	0	5	5	0.085	0.94	27.5
										15	35	0	5	6	0.059	0.93	30.6	0	5	6	0.059	0.93	30.6
										19	31	2	5	5	0.082	0.92	22.1	1	6	5	0.082	0.93	25.9
										19	35	2	5	6	0.064	0.92	23.1	1	6	6	0.055	0.93	28.1
0.20	0.40	22	44	5	8	12	0.095	0.90	26.6	20	42	4	8	12	0.071	0.90	27.4	3	9	12	0.063	0.91	32.7
										20	46	3	9	13	0.063	0.93	35.0	3	9	13	0.063	0.93	35.0
										24	42	5	9	12	0.072	0.91	29.5	4	10	12	0.064	0.91	33.5
										24	46	5	9	13	0.072	0.92	30.8	4	10	13	0.063	0.93	35.6
0.30	0.50	21	42	6	11	16	0.098	0.90	29.9	19	40	5	10	16	0.077	0.87	29.4	4	11	16	0.066	0.87	33.9
										19	44	4	11	17	0.083	0.91	36.7	4	11	17	0.083	0.91	36.7
										23	40	6	12	16	0.069	0.87	32.2	6	12	16	0.069	0.87	32.2
										23	44	6	12	17	0.086	0.91	34.3	6	12	17	0.086	0.91	34.3
0.40	0.60	24	47	10	14	23	0.098	0.90	30.8	22	45	9	13	23	0.080	0.86	29.4	8	14	22	0.091	0.91	34.1
										22	49	8	14	24	0.083	0.92	36.2	8	14	24	0.083	0.92	36.2
										26	45	10	16	22	0.089	0.91	34.7	10	16	22	0.089	0.91	34.7
										26	49	10	16	24	0.082	0.92	36.5	10	16	24	0.082	0.92	36.5
0.50	0.70	24	47	13	18	27	0.097	0.90	30.0	22	45	12	17	26	0.090	0.88	27.8	10	16	27	0.078	0.90	34.8
										22	49	10	16	29	0.086	0.93	37.1	10	16	29	0.086	0.93	37.1
										26	45	13	18	27	0.081	0.90	33.3	13	19	27	0.070	0.90	33.7
										26	49	13	18	29	0.088	0.93	34.8	13	19	29	0.077	0.92	35.4
0.60	0.80	19	38	12	16	26	0.098	0.90	24.4	17	36	10	15	25	0.089	0.90	25.3	9	15	25	0.093	0.91	28.9
										17	40	9	15	28	0.075	0.91	31.4	9	15	28	0.075	0.91	31.4
										21	36	12	17	26	0.064	0.85	28.3	12	18	25	0.091	0.91	28.7
										21	40	12	17	28	0.086	0.92	30.2	12	18	28	0.073	0.91	30.7
0.70	0.90	15	29	11	14	23	0.095	0.91	18.7	13	27	9	13	22	0.062	0.87	18.8	8	13	22	0.063	0.88	22.0
										13	31	8	13	25	0.067	0.92	24.6	8	13	25	0.067	0.92	24.6
										17	27	12	16	22	0.065	0.88	20.7	12	16	22	0.065	0.88	20.7
										17	31	12	16	25	0.069	0.92	22.2	12	16	25	0.069	0.92	22.2

Table 1b: Attained designs using type I and II error probability spending functions versus planned category II optimal designs ($\alpha \le 0.10$, $1-\beta \ge 0.90$).

the null hypothesis s versus the alternative hypothesis H_1: $p>p_1$=0.30. The planned design is (n_1, n_2, a, b, c)=(17, 16, 2, 5, 5) with a significance level of 0.084, a power of 0.91, and an average sample size under the null hypothesis of 20.5. The attained sample sizes are n_1^*=19 and n_2^*=12 (n^*=31) at stages 1 and 2, respectively. Using the type I and II error probability spending functions defined in (9) and (10), we obtain the attained design (n_1^*, n_2^*, a, b, c) =(19, 12, 2, 5, 5). This design has a significance level of 0.082, a power of 0.92, and an average sample size under the null hypothesis of 22.1. The corresponding Green and Dahlberg design is (n_1^*, n_2^*, a, b, c) =(19, 12, 1, 6, 5). This design has a significance level of 0.082, a power of 0.93, and an average sample size under the null hypothesis of 25.9. In all 32 cases we investigated (Table 1b), the average performance of the proposed designs on significance and power is almost identical to that of Green and Dahlberg (mean significance level: 0.074 vs. 0.073; mean power: 0.91 vs. 0.91). In fact, in 17 cases out of the 32, the proposed designs are the same as those of Green and Dahlberg. The agreement between the proposed designs and the designs of Green and Dahlberg is due to the fact that the type I and II error probabilities computed by (9) and (10) are close to 0.02 in many cases. In the remaining 15 cases, the proposed designs uniformly have smaller average sample size under the null hypothesis than those of Green and Dahlberg (mean average sample size: 27.5 vs. 31.0).

Proposed designs with adjustments by the end of stages 1 and 2 versus planned category I designs

Examples of proposed designs with adjustments by the end of stages 1 and 2 versus planned category I designs are presented in Table 2a. For example, at the 9th row entry, the testing procedure is for the null hypothesis H_0: $p \le p_0$=0.10 versus the alternative hypothesis H_1: $p>p_1$=0.30. The planned design is (n_1, n_2, a, c)=(17, 16, 2, 5) with a significance level of 0.081, a power of 0.90, and an average sample size under the null hypothesis of 20.8 (Table 1a). The attained sample size is n_1^* =19 at stage 1. For given n_1^* =19, we obtain the adjusted design (n_1^*, n_2^*, a^*, c^*)=(19, 11, 2, 5) satisfying the significance and power requirements and minimizing the average sample size under the null hypothesis. The target of accrual at stage 2 will be n_2^{**} =11 at stage 2. The attained sample size at stage 2 is n_2^{**} =9 (n^{**}=28). For given (n_1^*, n_2^{**}, a^*) =(19, 9, 2), we found that the threshold at stage 2 should be c^{**} = c^* =5. The attained design $(n_1^*, n_2^{**}, a^*, c^{**})$ =(19, 9, 2, 5) has a significance level of 0.052, a power of 0.88, and an average sample size of 21.7. The corresponding Green and Dahlberg design is $(n_1^*, n_2^{**}, a^*, c^{**})$ =(19, 9, 1, 5). This design has a significance level of 0.055, a power of 0.89, and an average sample size under the null hypothesis of 24.2. In all 32 cases we investigated (Table 2a), the average performance of the proposed

		Planned Sample Sizes		Actual Sample Size at Stage 1	Adjusted Design by the end of Stage 1						Actual total Sample Size	Attained Design								
												Proposed					Green & Dahlberg				
p_0	p_1	n_1	n	n_1^*	n^*	a^*	c^*	α^*	$1-\beta^*$	AVE N	n^{**}	a^*	c^{**}	α^*	$1-\beta^*$	AVE N	a^*	c^{**}	α^*	$1-\beta^*$	AVE N
0.05	0.20	19	38	17	33	0	3	0.078	0.91	26.3	31	0	3	0.065	0.89	25.1	0	3	0.065	0.89	25.1
				17							35	0	3	0.091	0.93	27.5	0	3	0.091	0.93	27.5
				21	34	1	3	0.078	0.90	24.7	32	1	3	0.068	0.89	24.1	0	3	0.073	0.91	28.3
				21							36	1	3	0.088	0.92	25.2	0	4	0.032	0.87	30.9
0.10	0.30	17	33	15	30	1	5	0.068	0.91	21.7	28	1	5	0.053	0.88	20.9	0	5	0.055	0.89	25.3
				15							32	1	5	0.086	0.93	22.7	0	5	0.093	0.95	28.5
				19	30	2	5	0.067	0.91	22.2	28	2	5	0.052	0.88	21.7	1	5	0.055	0.89	24.2
				19							32	2	5	0.083	0.92	22.8	1	5	0.092	0.95	26.5
0.20	0.40	20	40	18	37	3	10	0.098	0.91	27.5	35	3	10	0.072	0.88	26.5	2	10	0.074	0.89	30.4
				18							39	3	11	0.070	0.90	28.5	2	11	0.073	0.91	33.3
				22	38	5	10	0.099	0.90	26.3	36	5	10	0.078	0.88	25.7	4	10	0.086	0.90	28.4
				22							40	5	11	0.073	0.89	26.8	4	11	0.083	0.92	30.2
0.30	0.50	21	42	19	55	6	20	0.090	0.90	31.0	53	6	20	0.068	0.89	30.4	4	20	0.084	0.95	43.4
				19							57	6	21	0.078	0.90	31.7	4	21	0.100	0.96	46.3
				23	45	7	17	0.089	0.91	31.4	43	7	17	0.062	0.87	30.6	6	17	0.065	0.89	34.2
				23							47	7	18	0.075	0.90	32.2	6	18	0.081	0.92	36.4
0.40	0.60	25	49	23	55	10	26	0.082	0.90	32.1	53	10	25	0.087	0.90	31.6	8	26	0.068	0.93	41.3
				23							57	10	27	0.077	0.90	32.8	8	27	0.098	0.96	43.8
				27	49	12	23	0.099	0.91	32.5	47	12	23	0.070	0.88	32.0	10	23	0.081	0.92	37.8
				27							51	12	24	0.093	0.91	33.0	10	25	0.072	0.92	40.0
0.50	0.70	24	47	22	41	11	24	0.098	0.91	29.9	39	11	23	0.095	0.89	29.1	10	23	0.099	0.90	31.9
				22							43	11	26	0.060	0.87	30.7	10	26	0.062	0.88	34.3
				26	43	14	25	0.096	0.91	30.7	41	14	24	0.094	0.90	30.2	13	25	0.058	0.86	32.3
				26							45	14	26	0.098	0.91	31.3	13	27	0.066	0.90	34.0
0.60	0.80	20	39	18	38	11	26	0.096	0.91	25.5	36	11	25	0.083	0.89	24.7	10	25	0.089	0.91	28.1
				18							40	11	28	0.064	0.89	26.2	10	28	0.069	0.91	30.4
				22	38	14	26	0.094	0.91	26.6	36	14	25	0.082	0.89	26.1	13	25	0.089	0.91	28.4
				22							40	14	28	0.064	0.89	27.2	13	28	0.069	0.91	30.2
0.70	0.90	15	29	13	29	9	23	0.086	0.92	19.7	27	9	22	0.058	0.87	18.9	8	22	0.059	0.87	22.1
				13							31	9	25	0.059	0.90	20.6	8	25	0.062	0.92	24.8
				17	31	13	24	0.094	0.91	19.8	29	13	23	0.074	0.89	19.4	12	23	0.090	0.93	21.7
				17							33	13	26	0.070	0.90	20.2	12	26	0.088	0.95	23.2

Table 2a: Attained designs with adjustments by the end of stage 1 versus planned type I optimal designs (α ≤ 0.10, 1-β ≥ 0.90).

designs on significance and power is similar to that of Green and Dahlberg (mean significance level: 0.075 vs. 0.076; mean power: 0.89 vs. 0.91). Only in 2 cases out of the 32, the proposed designs are the same as those of Green and Dahlberg. In the remaining 30 cases, the proposed designs uniformly have smaller average sample size under the null hypothesis than those of Green and Dahlberg (mean average sample size: 26.8 vs. 31.7).

Proposed designs with adjustments by the end of stages 1 and 2 versus planned category II designs

Examples of proposed designs with adjustments by the end of stages 1 and 2 versus planned category II designs are presented in Table 2b. For example, at the 9th row entry, the testing procedure is for the null hypothesis H_0: $p \le p_0$=0.10 versus the alternative hypothesis H_1: $p>p_1$=0.30. The planned design is (n_1, n_2, a, b, c)=(17, 16, 2, 5, 5) with a significance level of 0.084, a power of 0.91, and an average sample size under the null hypothesis of 20.5 (Table 1b). The attained sample size is n_1^*=19 at stage 1. For given n_1^* =19, we obtain the adjusted design $(n_1^*, n_2^*, b^*, a^*, c^*)$ =(19, 11, 2, 5, 5) satisfying the significance and power requirements and minimizing the average sample size under the null hypothesis. The target of accrual at stage 2 will be n_2^{**}= 11 at stage 2. The attained sample size at stage 2 is n_2^{**} =9 (n^{**}= 28). For given $(n_1^*, n_2^{**}, a^*, b^*)$ =(19, 9, 2, 5), we found that the threshold at stage 2 should be $c^{**} = c^* = 5$. The attained design $(n_1^*, n_2^{**}, a^*, b^*, c^{**})$ = (19, 9, 2, 5, 5) has a significance level of 0.063, a power of 0.88, and an average sample size of 21.3. The corresponding Green and Dahlberg design is $(n_1^*, n_2^{**}, a^*, b^*, c^{**})$ =(19, 9, 1, 6, 5). This design has a significance level of 0.055, a power of 0.89, and an average sample size under the null hypothesis of 24.1. In all 32 cases we investigated (Table 2B), the average performance of the proposed designs on significance and power is similar to that of Green and Dahlberg (mean significance level: 0.081 vs. 0.075; mean power: 0.89 vs. 0.91). In all 32 cases, the proposed designs are different than those of Green and Dahlberg, and the proposed designs uniformly have smaller average sample sizes under the null hypothesis than that of Green and Dahlberg (mean average sample size: 26.2 vs. 30.8).

A Real Example

Investors want to evaluate the efficacy of lenalidomide in a phase II clinical trial in non-Hodgkin's lymphoma patients who responded and then relapsed after the first chemotherapy. The null and alternative hypotheses are specified in (1) with $p_{0=}$0.20 and p_1=0.40. The required

		Planned Sample Sizes		Actual Sample Size at Stage 1	Adjusted Design by the End of Stage 1							Actual Total Sample Size	Attained Design											
													Proposed Design						Green & Dahlberg					
p_0	p_1	n_1	n	n_1^*	n^*	a^*	b^*	c^*	α^*	$1-\beta^*$	AVE N	n^{**}	a^*	b^*	c^{**}	α^*	$1-\beta^*$	AVE N	a^*	b^*	c^{**}	α^*	$1-\beta^*$	AVE N
0.05	0.20	19	38	17	32	0	3	3	0.091	0.91	25.0	30	0	3	3	0.081	0.88	23.9	0	4	3	0.059	0.87	24.5
				17	32							34	0	3	4	0.061	0.87	26.0	0	4	3	0.084	0.92	26.7
				21	38	1	3	4	0.095	0.90	24.4	36	1	3	4	0.092	0.89	24.0	0	4	4	0.039	0.88	30.6
				21	38							40	1	3	4	0.098	0.91	24.8	0	4	4	0.053	0.92	33.1
0.10	0.30	17	33	15	29	1	4	5	0.088	0.90	20.5	27	1	4	5	0.077	0.88	19.7	0	5	5	0.050	0.87	24.4
				15	29							31	1	4	6	0.069	0.88	21.3	0	5	5	0.085	0.94	27.5
				19	30	2	5	5	0.075	0.91	21.9	28	2	5	5	0.063	0.88	21.3	1	6	5	0.055	0.89	24.1
				19	30							32	2	5	5	0.089	0.93	22.4	1	6	5	0.092	0.95	26.4
0.20	0.40	22	44	20	37	4	9	10	0.095	0.90	26.1	35	4	9	10	0.071	0.87	25.4	3	9	10	0.075	0.89	28.7
				20	37							39	4	9	11	0.069	0.89	26.8	3	9	11	0.075	0.91	31.0
				24	36	5	9	10	0.091	0.90	27.7	34	5	9	10	0.070	0.87	27.1	4	10	10	0.063	0.86	29.3
				24	26							38	5	9	11	0.072	0.89	28.3	4	10	11	0.064	0.89	31.4
0.30	0.50	21	42	19	57	6	10	21	0.093	0.90	30.5	55	6	10	21	0.076	0.89	29.9	4	11	21	0.076	0.94	44.5
				19	57							59	6	10	22	0.084	0.90	31.1	4	11	22	0.090	0.96	47.3
				23	42	7	12	16	0.093	0.90	29.8	40	7	12	16	0.067	0.86	29.1	6	12	16	0.069	0.87	32.2
				23	42							44	7	12	17	0.081	0.90	30.6	6	12	17	0.086	0.91	34.3
0.40	0.60	24	47	22	49	9	13	24	0.098	0.90	30.7	47	9	13	24	0.079	0.87	30.0	8	14	23	0.087	0.91	35.1
				22	49							51	9	13	25	0.096	0.90	31.3	8	14	25	0.080	0.92	37.2
				26	45	11	15	22	0.099	0.90	31.2	43	11	15	22	0.076	0.86	30.7	10	16	21	0.094	0.91	33.8
				26	45							47	11	15	23	0.096	0.90	31.8	10	16	23	0.085	0.91	35.6
0.50	0.70	24	47	22	41	11	17	24	0.100	0.91	29.7	39	11	17	23	0.096	0.89	28.9	10	16	24	0.065	0.85	31.5
				22	41							43	11	17	26	0.063	0.87	30.6	10	16	26	0.074	0.89	33.7
				26	43	14	19	25	0.098	0.91	30.5	41	14	19	24	0.095	0.90	30.0	13	19	25	0.062	0.86	32.1
				26	43							45	14	19	26	0.100	0.91	31.0	13	19	17	0.070	0.90	33.8
0.60	0.80	19	38	17	40	10	14	28	0.092	0.90	26.2	38	10	14	27	0.082	0.88	25.4	9	15	27	0.062	0.88	30.2
				17	40							42	10	14	30	0.074	0.88	27.0	9	15	29	0.089	0.94	32.7
				21	36	13	17	25	0.097	0.90	25.7	34	13	17	24	0.084	0.88	25.0	12	18	24	0.075	0.88	27.7
				21	36							38	13	17	27	0.073	0.88	26.3	12	18	27	0.060	0.88	29.7
0.70	0.90	15	29	13	31	9	12	25	0.100	0.92	19.4	29	9	12	24	0.083	0.87	18.7	8	13	23	0.096	0.94	23.3
				13	31							33	9	12	27	0.086	0.90	20.1	8	13	26	0.097	0.96	25.9
				17	35	13	15	28	0.098	0.90	19.2	33	13	15	27	0.090	0.88	19.0	12	16	26	0.095	0.95	22.9
				17	35							37	13	15	30	0.091	0.90	19.4	12	16	29	0.093	0.96	24.4

Table 2b: Attained designs with adjustments by the end of stage 1 versus planned type II optimal designs (α ≤ 0.10, 1-β ≥ 0.90).

significance level and power are $\alpha \leq 0.10$ and 1- ≥ 0.90, respectively. The investigators planned to use the category I design with equal or almost equal sample sizes between two stages and with the average sample size under the null hypothesis minimized. The planned design is (n_1, n_2, a, c)=(20, 20, 4, 11).This design has a significance level of 0.078, a power of 0.90, and an average sample size under the null hypothesis of 27.4 (Table 1a).

Assume that the attained sample sizes are 18 and 20 in stages 1 and 2 (n^*=38), respectively. Using the type II error probability spending function defined in (9), we obtain the attained design (n_1^*, n_2^*, a^*, c^*) =(18, 20, 3, 11) (Table 1a). This design has a significance level of 0.060, a power of 0.88, and an average sample size under the null hypothesis of 28.0 (Table 1a).

By the method with redesigns adjusted to the attained sample sizes, we first obtain the adjusted design (n_1^*, n_2^*, a, c)=(18, 19, 3, 10) given the attained sample size of n_1^*=18 at the first stage (Table 2a). This design satisfies the significance and power requirements and minimizes the average sample size under the null hypothesis. The target accrual at stage 2 is 19 patients now. Assume that the attained sample size is n_2^{**}=17 at the second stage (n^{**}=35). We found that the threshold at the second stage $c^{**} = c^* = 10$ satisfies the significance level requirement. The final attained design is $(n_1^*, n_2^{**}, a^*, c^{**})$=(18, 17, 3, 10) (Table 2a). This design has a significance level of 0.072, a power of 0.88, and an average sample size of 26.5 (Table 2a).

Discussion and Conclusion

We have proposed two alternative designs to replace planned two-stage designs when the attained sample sizes at stage 1 and 2 are different than the planned sample sizes. The objectives of the proposed designs are not only to satisfy the significance level requirement and approximately satisfy the power requirement, but also to be close to the planned design in terms of the original criteria, such as minimizing the average sample size.

In our first approach, we define type I and type I error probability spending functions by the planned design, and then we conduct two-stage testing using these error probability spending functions. In contrast, Green and Dahlberg proposed to spend type II error probability at stage 1 at a fixed level of 0.02. Since the planned design is used in determination of error probabilities spent at the first stage, the proposed designs are closer to the planned designs than those of Green

and Dahlberg. After the planned design is set up, the error probability spending function can be specified. Therefore, both the planned design and the alternative design can be specified in the protocol before the study starts.

In our second approach, we generate a modified design using the same criteria as in the planned design, conditional on the attained sample size at the first stage. A new accrual target is set up for the second stage according to the modified design. When the attained sample size is different than the redesigned sample size at stage 2, another adjustment for the threshold at stage 2 may be needed to satisfy the significance requirement. The dynamic redesigns depend only on attained sample sizes, and are independent of response data. The strategy of the testing procedure can be specified in the protocol before the study starts by tabulating some possible modified designs.

In our numerical investigations, we considered planned designs that minimize the average sample size under the null hypothesis, and we found that the proposed alternative designs had smaller average sample size under the null hypothesis than those of Green and Dahlberg.

Our program was written in PROC IML, SAS software version 9.3. The program is available at <http://users.phhp.ufl.edu/mchang/attained versus planned design/>>.

References

1. Chang MN, Therneau TM, Wieand HS, Cha SS (1987) Designs for group sequential phase II clinical trials. Biometrics 43: 865-874.
2. Chen TT (1997) Optimal three-stage designs for phase II cancer clinical trials. Statistics in Medicine 16: 2701-2711.
3. Ensign LG, Gehan EA, Kamen DS, Thall PF (1994) An optimal three-stage design for phase II clinical trials. Statistics in Medicine 13: 1727-1736.
4. Jung SH, Carey M, Kim KM (2001) Graphical search for two-stage designs for phase II clinical trials. Controlled Clinical Trials 22: 367-372.
5. Shuster J (2002) Optimal two-stage designs for single arm phase II cancer trials. Journal of Biopharmaceutical Statistics 12: 39-51.
6. Simon R (1989) Optimal two-stage designs for phase II clinical trials. Controlled Clinical Trials 10: 1-10.
7. Therneau TM, Wieand HS, Chang MN (1990) Optimal designs for a grouped sequential binomial trial. Biometrics 46(3): 771-783.
8. Green SJ, Dahlberg S (1992) planned versus attained design in phase II clinical trials. Statistics in Medicine 11: 653-862.
9. Herndon JE (1998) A design alternative for two-stage, phase II, multicenter cancer clinical trials. Controlled clinical trials 19: 440-450.
10. Matsumoto K, Katsumata N, Saito I, Shibata T, Konishi I, et al. (2012) Phase II study of oral etoposide and intraveneous irinotecan for patients with platinum-resistant and taxane-pretreated ovarian cancer: Japan Clinical Oncology Group Study 0503. Japanese Journal of Clinical Oncology 42: 222-225.
11. Lan KKG, Demets DL (1983) Discrete sequential boundaries for clinical trials. Biometrika 70: 659-663.
12. Chang MN, Hwang IK, Shih WJ (1998) Group sequential designs using both type I and type II error probability spending functions. Communication in Statistics, Theory and Methods 27: 1323-1339.
13. Hampson LV and Jennison C (2013) Group sequential tests for delayed responses. JRSS B 75: 3-54.