Published in final edited form as:

Stat Med. 2012 July 20; 31(16): 1688–1698. doi:10.1002/sim.5314.

Improved Two-Stage Tests for Stratified Phase II Cancer Clinical Trials

Myron N. Chang¹, Jonathan J. Shuster², and Wei Hou¹

¹Department of Biostatistics, University of Florida, Gainesville, FL, U.S.A

²Department of Health Outcomes and Policy, University of Florida, Gainesville, FL, U.S.A

Summary

In a single-arm, two-stage, phase II cancer clinical trial for efficacy screening of cytotoxic agents, a common primary endpoint is a binary (yes/no) patient response to treatment. Usually, fixed decision boundaries are used in binomial tests to determine whether the study treatment is promising enough to be studied in a large-scale, randomized phase III trial. We may know in advance that the patient response distribution for a phase II clinical trial will be heterogeneous, making it advisable to stratify patients into subgroups, each with a different prognosis. In this case, fixed decision boundaries may be inappropriate. In this article, we propose two-stage tests based on the Neyman-Pearson lemma. The proposed test statistic is a linear combination of the observed number of responders in each stratum. The test allows adjustment of the decision boundaries to the observed numbers of patients in each stratum and permits sample sizes to be increased adaptively after the originally planned number of patients is observed at each of the two stages. Our numerical results show that the proposed test is more powerful than an existing test in many cases. Finally, we present an application to a Children's Oncology Group (COG) phase II clinical trial in patients who relapsed after initial treatment for neuroblastoma.

Keywords

Adaptive design; Decision boundaries; Neyman-Pearson lemma; Likelihood ratio; Type I and type II error probability spending functions

1. INTRODUCTION

Phase II cancer clinical trials, especially for rare diseases, are often single-arm studies conducted to decide whether an experimental treatment is sufficiently promising to be studied in a larger-scale phase III randomized trial. The endpoint of a phase II clinical trial is typically a binary patient response to the study treatment. For both ethical and efficiency reasons, most such trials use sequential designs. In practice, clinical trials usually are conducted as multistage experiments, rather than being fully sequential. Two-stage designs are commonly used for phase II cancer clinical trials because of their logistical simplicity and because the benefits of multistage trials diminish after two stages. Many approaches to the design of multistage phase II clinical trials have been proposed [1-7].

The patient population for a phase II cancer clinical trial is sometimes heterogeneous. Patients frequently can be stratified according to age, gender, disease stage, and/or other risk factors. Binomial tests in these trials can be highly inefficient because such tests ignore the heterogeneity of response rates across strata. Due to the small sample sizes typically available for phase II cancer clinical trials, it is also inefficient to conduct independent binomial tests within strata because the totality of information will be lost for primary inferential purposes. London and Chang [8] have proposed unconditional and conditional

designs. Their conditional design allows decision boundaries to be changed depending on the observed numbers of patients in the defined strata. Their test is based on the difference between the observed total number of responders across strata and the expected total number of responders under the null hypothesis. The method proposed in this article is an extension of their conditional method for a single-arm trial. Hereafter, when we refer to the design by London and Chang [8], we will be referring to the conditional design. Other authors have proposed designs of phase II clinical trials with heterogeneous patient populations [9-13]. The key difference between the proposed method and the methods in [9-13] will be described in Discussion section.

A drawback of the test proposed by London and Chang [8] is that equal weights are used for all outcomes in the construction of the test statistic. This approach leads to suboptimal tests in many cases. Another drawback is that the significance levels of the tests are often smaller than the nominal levels due to the discreteness of the distribution of the test statistic. In this article, we propose a test based on the Neyman-Pearson lemma [14] and conditional on the numbers of patients who were enrolled into each stratum. The test statistic is the ratio of the likelihood under the alternative hypothesis to that under the null hypothesis. We found that the test statistic is equivalent to a linear combination of the numbers of responders in the strata. The coefficient corresponding to a stratum is the log of the odds ratio of the response rate under the alternative hypothesis to that under the null hypothesis in the same stratum. Two-stage tests are proposed based on exact computations. Our numerical studies show that the proposed tests have significance levels close to the nominal levels and are more powerful than the tests proposed by London and Chang [8] in many cases. Since the true stratum-specific target population proportions of responders for an effective treatment may differ from those specified under the actual alternative hypothesis, we also numerically study the robustness of the performance of the proposed testing procedures. Finally, we apply the proposed testing procedure to a phase II COG study in relapsed neuroblastoma patients.

The remainder of the article is organized in five sections. In Section 2, we derive the proposed tests and define the two-stage testing procedures. We present numerical examples in Section 3, and sample-size determination and guidance for the application in Section 4. In Section 5, we apply the proposed method to a real study, and in Section 6, we discuss the key difference between the proposed method and other methods.

2. TWO-STAGE TESTS

The objective of a phase II clinical trial is to evaluate an experimental treatment that potentially will increase response rates over an historical baseline. To apply the proposed method, we require prior knowledge of the existence of subgroups (strata), stratum specific null hypothesized response rates based on past history and the projected stratum specific improvement in response. Assume that patients are stratified into k strata. Let p_i denote the expected response rate of the study treatment, and let p_{i0} denote the historical response rate for patients in stratum i , $i = 1, \dots, k$. We are interested in testing the null hypothesis $H_0: p_i \leq p_{i0}$, $i = 1, 2, \dots, k$, vs. the alternative hypothesis $H_a: p_i > p_{i0}$, $i = 1, 2, \dots, k$, with a desired significance level α and power $(1 - \beta)$. The power will be evaluated at $p_{i1} = p_{i0} + \Delta_i$, $i = 1, 2, \dots, k$, where Δ_i is the specified improvement in response rate clinicians are expecting from the study treatment.

Let R_i be the number of responders and N_i be the number of patients enrolled in stratum i , $N = \sum_{i=1}^k N_i$, and $R = \sum_{i=1}^k R_i$. The tests proposed by London and Chang [8] are based on the total number of responders R , conditional on the observed number of patients $N_i = n_i$ in stratum i , $i = 1, \dots, k$. Exact computations are used to obtain decision boundaries. For given

$N_i = n_i$, the R_i 's are independent binomial random variables with $R_i \sim \text{binomial}(n_i, p_i)$. Denote $\mathbf{p} = (p_1, p_2, \dots, p_k)$, $\mathbf{p}_0 = (p_{10}, p_{20}, \dots, p_{k0})$, and $\mathbf{p}_1 = (p_{11}, p_{21}, \dots, p_{k1})$. The likelihood function for $R_i = r_i$, $i = 1, 2, \dots, k$, conditional on $N_i = n_i$, $i = 1, 2, \dots, k$, is:

$$L(\mathbf{p}) = \prod_{i=1}^k \binom{n_i}{r_i} p_i^{r_i} (1 - p_i)^{n_i - r_i}.$$

In this article, we will propose the most powerful test based upon the Neyman-Pearson lemma. The test statistic is the log of the ratio of the likelihood under parameters $\mathbf{p}_1$ to that under parameters $\mathbf{p}_0$:

$$\sum_{i=1}^k \left(r_i \log \frac{p_{i1}(1 - p_{i0})}{p_{i0}(1 - p_{i1})} + n_i \log \frac{1 - p_{i1}}{1 - p_{i0}} \right).$$

The above test statistic is equivalent to

$$T = \sum_{i=1}^k w_i R_i,$$

where

$$w_i = \log \frac{p_{i1}(1 - p_{i0})}{p_{i0}(1 - p_{i1})}. \quad (1)$$

If all w_i 's are equal (common odds ratios), then the proposed test is the same as in London and Chang [8]. If the w_i 's are unequal, then the proposed test should tend to be more powerful than that in London and Chang [8]. If it is ever less powerful, it is strictly due to discreteness with the true rejection probability of London and Chang [8] attaining a closer approximation to the nominal value than this method based upon Neyman-Pearson theory.

Assume that a two-stage design for testing H_0 vs. H_a uses M_1 patients at the first stage and M_2 patients at the second stage. Let n_{ij} be the number of patients entered in stratum i at stage j , and let R_{ij} be the number of responders among the n_{ij} patients, $i = 1, 2, \dots, k$, and $j = 1, 2$. The two-stage testing procedure works as follows: When M_1 patients are enrolled at the first stage, and when the response data on the M_1 patients are available, the test statistic will be computed:

$$T_1 = \sum_{i=1}^k w_i R_{i1},$$

where w_i is as in (1). If $T_1 < a$, then we declare that the study treatment is not promising and the study is stopped; if $T_1 > b$, then we claim that the study treatment is promising and the study also is stopped; if $a \leq T_1 \leq b$, then the accrual will continue to the second stage, where decision boundaries a and b are chosen depending on the observed numbers of patients, n_{i1} , in stratum i at the first stage. When the response data on the M_2 patients from the second stage are available, then the following test statistic will be computed:

$$T_2 = \sum_{i=1}^k w_i (R_{i1} + R_{i2}).$$

If $T_2 > c$, then we claim that the study treatment is promising; otherwise, we will conclude that the treatment is not promising. The decision boundary c is chosen depending on the observed numbers of patients, n_{ij} , in stratum i at both stages, $i = 1, 2, \dots, k$.

We next address the problem of determination of the decision boundaries a , b , and c for given n_{ij} 's. We propose to set up, before the study starts, a fixed fraction γ_1 of type I error 1 probability and a fixed fraction γ_2 of type II error probability that will be spent at the first stage [15]. The decision boundary a is chosen as the largest real number satisfying

$$P(T_1 < a | n_{i1}, p_{i1}, i=1, 2, \dots, k) \leq \gamma_2 \beta. \quad (2)$$

The decision boundary b is chosen as the smallest real number satisfying

$$P(T_1 > b | n_{i1}, p_{i1}, i=1, 2, \dots, k) \leq \gamma_1 \alpha. \quad (3)$$

The decision boundary c is chosen as the largest real number satisfying

$$P(T_1 > b | n_{i1}, p_{i1}, i=1, 2, \dots, k) + P(a \leq T_1 \leq b, T_2 > c | n_{ij}, p_{ij}, i=1, 2, \dots, k, j=1, 2) \leq \alpha. \quad (4)$$

The power of the test is

$$\text{Power} = P(T_1 > b | n_{i1}, p_{i1}, i=1, 2, \dots, k) + P(a \leq T_1 \leq b, T_2 > c | n_{ij}, p_{ij}, i=1, 2, \dots, k, j=1, 2). \quad (5)$$

The design with decision boundaries a , b , and c guarantees that the significance level does not exceed α

3. Examples

Examples with three strata are listed in Table 1. We are testing the null hypothesis $H_0: \mathbf{p} = \mathbf{p}_0 = (p_{10}, p_{20}, p_{30})$, and the power is calculated under $H_a: \mathbf{p} = \mathbf{p}_1 = (p_{11} + \Delta_1, p_{21} + \Delta_2, p_{31} + \Delta_3)$. We use $\Delta_1 = \Delta_2 = \Delta_3 = 0.2$ in all examples in Table 1. Note that the equal Δ_j 's are just for convenience of tabulation. Our computer program allows for unequal treatment effects. The significance level is taken as 0.05. The weights w_i are computed by formula (1). For example, if $\mathbf{p}_0 = (0.5, 0.3, 0.1)$ and $\mathbf{p}_1 = (0.7, 0.5, 0.3)$, then $w_1 = 0.847$, $w_2 = 0.847$ and $w_3 = 1.350$, respectively, for the three strata (first 17 entries of Table 1). The test statistics at stages 1 and 2 are T_1 and T_2 as in (2) and (3), respectively. A type I error probability of 0.0125 ($\gamma_1 = 0.25$) and a type II error probability of 0.05 ($\Delta_2 = 0.25$) will be spent at the first stage. It is assumed that the sample sizes at stages 1 and 2 are equal ($M_1 = M_2$), and the number of patients in each stratum at stages 1 and 2 are also equal ($n_{i1} = n_{i2}$). Note that the equal sample sizes are just for convenience of tabulation. Our computer program generates designs for equal and unequal sample sizes as well. The proposed two-stage testing procedure is defined in the previous section using the decision boundaries a , b , and c . The test statistics proposed by London and Chang [8] are centered by the expected response under the null hypothesis. Since the expected response under the null hypothesis is constant

when the test is conditional on the observed sample sizes in the strata, their test statistics are equivalent to

$$T_1 = \sum_{i=1}^3 R_{i1} \quad \text{and} \quad T_2 = \sum_{i=1}^3 (R_{i1} + R_{i2}),$$

at stages 1 and 2, respectively. We obtained the decision boundaries, a , b , c , the significance level, α , and the power, $(1 - \beta)$, in Table 1 by exact computation.

From the examples in Table 1, it can be seen that the power of the proposed test, with few exceptions, is higher than that of the test proposed by London and Chang [8]. The significance levels of the proposed test, again with few exceptions, are closer to the nominal level (0.05) than are the significance levels of the test proposed by London and Chang [8].

It will be difficult in practice to identify the alternative hypothesis exactly, in particular at earlier stages of the drug development process. When the real treatment effect does not match the anticipated improvement as indicated in the alternative hypothesis before the study starts, the weights w_j 's would be suboptimal. Therefore, it is important to study the power performance when the true response rates differ from those specified under the alternative hypothesis. Some examples are presented in Table 1. The designs are obtained for testing the null hypothesis $H_0: \mathbf{p} = \mathbf{p}_0 = (p_{10}, p_{20}, p_{30})$ vs. the alternative hypothesis $H_a: \mathbf{p} = \mathbf{p}_1 = (p_{10} + 0.2, p_{20} + 0.2, p_{30} + 0.2)$, and the weights w_j 's are computed by (1) using parameters $\mathbf{p}_0$ and $\mathbf{p}_1$ as described before. Assume that the true response rates $\mathbf{p} = \mathbf{p}_2 = (p_{10} + 0.1, p_{20} + 0.25, p_{30} + 0.25)$, denoted as H_a^* . The power of the proposed testing procedure and that of the test proposed by London and Chang [8] under H_a^* are listed in Table 1. It can be seen from Table 1 that the power of the proposed testing procedure to detect the treatment effect in terms of response rate is still generally higher than the power of the test proposed by London and Chang [8].

4. SAMPLE-SIZE DETERMINATION AND GUIDANCE FOR APPLICATION

Denote P_i as the true proportion of patients in stratum i , $i = 1, 2, \dots, k$. Note that $\sum_{i=1}^k P_i = 1$. To determine the sample sizes M_1 and M_2 , a rough estimate of P_i for each stratum is required. The sample sizes will be determined before the study starts as follows.

1. Set up an initial value for $M_1 = M_2$ that is smaller than the anticipated sample size. For example, $M_1 = M_2 = 10$.
2. Set n_{i1} and n_{i2} to be the integer closest to $M_j P_i$, $i = 1, 2, \dots, k$, $j = 1, 2$, under the constrain $\sum_{i=1}^k n_{ij} = M_j$, $j = 1, 2$.
3. Determine the decision boundaries a , b , and c by (2) – (4).
4. Compute the power using the decision boundaries a , b , and c by (5).
5. If the power is less than $1 - \beta$, then increase M_1 by 1 when $M_1 = M_2$ and increase M_2 by 1 when $M_1 > M_2$.
6. Repeat steps 2 – 5 until the power requirement is satisfied.

Based on initial estimates of proportions of patients in strata, and assuming $n_{i1} = n_{i2}$, $i = 1, 2, \dots, k$, one determines the sample sizes M_1 and M_2 as above. Although one can control the n_{ij} by temporary closures of strata when they meet their requirement, this could lengthen the accrual process substantially and deny otherwise eligible subjects access to the treatment. In

practice, accrual in each stratum is usually left open until the total accrual for the stage is completed, regardless of the actual stratum distribution. Therefore, due to the randomness of the numbers of patients who enter the k strata, the observed number of patients, n_{ij} , may not be close to the anticipated number, $M_j P_i$, and hence, the power may be below the intended level. We propose that after M_1 patients enter the study and the number of patients, n_{i1} , in stratum i is observed, $i = 1, 2, \dots, k$, at the first stage, the decision boundaries a , b , and c are determined under the assumption that $n_{i1} = n_{i2}$, $i = 1, 2, \dots, k$, and the power is assessed to see whether the sample sizes are sufficient to satisfy the power requirement. At this time point, if the projected power is deficient, the sample sizes will be increased to achieve the required power. One would not need to close the study temporarily to wait for outcome data, since the power re-assessment is based only on stratum sample sizes, which are immediately observable. Hence, this reassessment can be done very rapidly. Indeed, the power reassessment can be planned for even before the first M_1 patients have entered the study. After M_2 patients enter the study and the number of patients, n_{i2} , in stratum i is observed, $i = 1, 2, \dots, k$, at the second stage, the decision boundary c will be determined and the power will be recomputed. If the power is lower than the required level, then we would increase M_2 ($M_2 > M_1$) and adjust the decision boundary c . When dynamic redesign is used, one should not look at the response data until the redesign is completed. Once these adjusted sample sizes are determined, and additional patients seem needed, we should look at the responses to date to see whether the inference can possibly be overturned by these additional patients; i.e., whether the current test statistic without additional patients exceeds the adjusted decision boundary c . If a definitive decision is assured, the study can be terminated without this dynamic adjustment. Based on our experience, sample-size adjustments are usually either not needed or needed once at the first stage. An alternative option to this adaptive design is to plan the sample size so that there is adequate power up front for a range of potential observed stratum sample sizes. This would require upfront effort, and may require larger sample sizes, but would not need frequent monitoring.

5. A REAL EXAMPLE

Due to the fact that most phase II oncology clinical trials take several years to accrue sufficient patients, we are unable to apply the proposed testing procedure to a prospective study. We retrospectively applied the proposed testing procedure to a phase II clinical trial in patients who relapsed after initial treatment for neuroblastoma, COG study P9462 [8,16]. The trial was designed to evaluate the combination therapy Topotecan + Cytosine. The study treatment is hypothesized to increase the response rate over historical expectations. It is well established that younger patients have a better response rate than older ones. Patients were stratified into three age groups: 1 year, 1-4 years, and 5 years. The hypothesized response rates under the null are $p_{10} = 0.35$, $p_{20} = 0.20$ and $p_{30} = 0.15$ for the three strata, respectively. The required type I error rate is 0.05 and the required power is 0.80 to detect a 20 percent increase in the response rate in each of the three strata. The weights in (1) are

$$w_1 = \log \frac{(0.55)(0.65)}{(0.35)(0.45)} = 0.820,$$

$$w_2 = \log \frac{(0.40)(0.80)}{(0.20)(0.60)} = 0.981, \quad \text{and}$$

$$w_3 = \log \frac{(0.35)(0.85)}{(0.15)(0.65)} = 1.116$$

From previous experience, the anticipated proportions of patients in the three strata are 0.10, 0.60 and 0.30, respectively. Using the anticipated patient proportions, P_i 's, and the above steps for sample-size determination, we found that sample sizes $M_1 = M_2 = 16$ would be sufficient to achieve the significance level and power requirements with the anticipated numbers of patients $n_{11} = n_{12} = 2$, $n_{21} = n_{22} = 9$, and $n_{31} = n_{32} = 5$. Following the design, 16 patients would enter the study at the first stage. The actual numbers of patients who entered the study were $n_{11} = 1$, $n_{21} = 12$, and $n_{31} = 3$. At this time point, the decision boundaries a , b , and c and the power would be calculated assuming $n_{i2} = n_{i1}$, $i = 1, 2, 3$. By the method developed in Section 2, we found that $a = 3.077$, $b = 6.974$, and $c = 10.186$, and the significance level $\alpha = 0.048$ and the power $(1-\beta) = 0.804$. Since the power requirement was achieved, no increase would have been needed for sample size M_1 . Note that if the power was < 0.80 , then the sample sizes $M_1 = M_2$ would have been increased and the decision boundaries and power would have been recalculated after the number of patients, n_{i1} , were observed among the first M_1 patients who entered the study. The actual numbers of responders among the first 16 patients in the three strata were $R_{11} = 1$, $R_{21} = 2$, and $R_{31} = 1$, respectively. Therefore, the observed value of the test statistic at the first stage would have been

$$T_1 = (0.820)(1) + (0.981)(2) + (1.16)(1) = 3.898.$$

Since $a < T_1 < b$, the study treatment would pass the first-stage test and the accrual would be continued to the second stage. After another 16 patients entered the study at the second stage, we found that the observed numbers of patients in the three strata at the second stage were $n_{12} = 1$, $n_{22} = 8$, and $n_{32} = 7$. For $a = 3.077$, $b = 6.974$, and conditional on the observed n_{ij} , we found that $c = 10.05$, significance level $\alpha = 0.048$, and power $(1-\beta) = 0.804$. Since the power requirement would have been achieved, no increase for the sample size M_2 would have been needed. Note that if the power had been below 0.80, then M_2 would have been increased ($M_2 > M_1$) and the decision boundary c and the power would have been recalculated after the actual numbers of patients, n_{i2} , were observed. In the actual study, $R_{12} = 0$, $R_{22} = 4$, and $R_{32} = 1$ for the 16 patients who entered the study at the second stage. Correspondingly, the test statistic at the second stage would have been

$$T_2 = (0.820)(1) + (0.981)(6) + (1.116)(2) = 8.938.$$

Since $T_2 < c$, we would have failed to reject the null hypothesis. Note that the required total sample size was 36 in London and Chang [8]. An 11.1% saving on sample size is achieved by the proposed test.

6. DISCUSSION

Many authors [8-13] have proposed designs for stratified phase II clinical trials. Jung, Chang, and Kang [9] proposed designs similar to London and Chang [8]. In their design, there is no upper decision boundary at the first stage, so even if the treatment is very effective as shown at the first stage, the study will not be stopped as long as the observed responses exceed a threshold at the first stage. They used $[\sum_{i=1}^k n_{i1} p_{i0}]$ as the lower decision boundary at the first stage, where $[x]$ represents the largest integer $\leq x$. Therefore,

they do not need to specify the spending of type I error and type II error probability at the first stage. The attractive feature of their designs is the simplicity and the ease of programming. Note that the designs proposed in this article include designs without an upper decision boundary at the first stage as special cases when the spending of type I error probability is set up as zero ($\gamma_1 = 0$). Practitioners can design their studies either with or without an upper decision boundary at the first stage.

Spoto and Gaynon [10] proposed tests based on the sum of responders in all strata conditional on the sample-weighted average response under the null hypothesis. The distribution of their test statistic still depends on the true proportions of patients in the strata. They derive the decision boundaries by minimizing an ad hoc objective function based on large-sample approximation. They showed by numerical investigation that their designs are robust for various proportions of patients in the strata. In their design, all possible decision boundaries for stages 1 and 2 are computed at the outset, and no recalibration of the design is necessary during the trial. In contrast, the proposed testing procedure is conditional on the observed sample sizes in the strata. Thus, the distribution of the test statistic is independent of the proportions of patients in the strata. We derive the decision boundaries by exact computation. Since typical phase II clinical trials have small sample sizes, a testing procedure based on normal approximation may not control error rates accurately. As pointed out in Section 3, an alternative to the proposed adaptive design is to plan the sample size so that there is adequate power up front for a range of potential observed stratum sample sizes. This would require upfront effort, and may require a larger sample size, but would not require recalibration during the trial.

Thall et al. [11] and Wathen et al. [12] proposed a Bayesian approach for the design of a stratified phase II clinical trial. They used physician input and information from previous studies to set up the prior joint distribution of response rates in the strata. Many model assumptions are needed in establishing the prior distribution. They used the posterior marginal distribution of the response rate in a stratum to determine whether the treatment is promising within the stratum. In contrast, we attempt to rely as little as possible on subjective input from physician or previous studies. In our design, the target is to evaluate the treatment effect across all strata, rather than within an individual stratum. Recently, Chang, Jung, and Wu [13] proposed frequentist designs including futility tests within stratum and global tests across all strata.

As an extension of designs by London and Chang [8], we have proposed tests based on the Neyman-Pearson lemma. If the weights, w_i , are equal, which will occur if and only if the alternative hypothesis has common odds ratios to the null hypothesis, then the proposed test is the same as the one proposed in London and Chang [8]. If the weights are unequal, then we expect that the proposed test will have higher power. The proposed tests have demonstrated improvement in power compared with the tests proposed by London and Chang [8] in most of our examples. The average power of the proposed test in the 68 examples in Table 1 was 0.86 compared with an average power of 0.81 for the test proposed by London and Chang [8]. The improvement in power is roughly equivalent to a 16% saving in sample size.

The weights w_i used in the Neyman-Pearson test statistic are real numbers instead of integers. Therefore, the sample space of the proposed test statistic is finer (less “lumpy”) than that proposed by London and Chang [8], whose outcome sample space is integer-based. The reduced discreteness of the distribution of the test statistic results in actual significance levels closer to the nominal level, and therefore, results in higher power. This is a desirable feature for clinical trials with relatively small sample sizes.

Note that the improvement in power from the proposed method is not uniform. For example, the row entry with $(p_{10}, p_{20}, p_{30}, n_{11}, n_{21}, n_{31}) = (0.5, 0.3, 0.1, 12, 4, 2)$ in Table 1 shows that the power of the test proposed by London and Chang [8] is slightly higher than the power of the currently proposed test (0.809 vs. 0.795). This is not a contradiction to the most powerful property of the Neyman-Pearson lemma. In the proposed testing procedure, two-stage tests are performed and type I and type II error probability spending functions are used. The proposed approach may destroy the theoretical property of the most powerful test. However, in 64 of the 68 examples we investigated, the power of the proposed test was superior to that of the test proposed by London and Chang [8]. For the four cases where power was lower, the differences in power were small.

As a caution, if dynamic redesign is used to account for the fact that the stratum-by-stage sample sizes, n_{ij} , are not exactly as planned, one should not look at the outcome data until the redesign is completed. Then and only then should the investigators determine whether a definitive decision can be made; i.e., whether the current test statistic at the second stage without additional patients exceeds the upper decision boundary c in the redesign, prior to enrolling additional patients into the stage. For sample-size recalculation, one would not need to pause the study to wait for outcome data since the power re-assessment is based only on stratum sample sizes, which are immediately observable, so this reassessment can be done very rapidly. In practice, the sample-size adjustment is usually either not needed or needed once at the first stage.

It is, of course, impossible to identify exactly the true target population proportion of responders in a stratum before the study starts. The alternative hypothesis is merely the treatment effect anticipated by clinicians. However, the weights w_j for the proposed testing procedure are computed under the alternative hypothesis as set up front. From our numerical studies, if the true treatment effects are not far from the alternative hypothesis, the power is still satisfactory in most cases. We suggest that the power performance of the proposed testing procedure with suboptimal weights be investigated for a range of plausible alternative hypotheses before the study starts. In designing a trial that uses this methodology, it is important to prospectively define strata, each with sufficient accrual potential to avoid empty cells by the end of the first stage.

All computations were exact, using independent binomial random variables. The complexity of the computations is similar to that in London and Chang [8]. The application program has been developed and implemented in the SAS IML language and is available from the first author upon request by email (mchang@biostat.ufl.edu).

Acknowledgments

This work was partially supported by grants 1UL1RR029890 and U54RR025208 from the National Institute of Research Resources, National Institutes of Health. We thank the reviewers and the associate editor for their constructive comments, which helped to improve this article.

REFERENCES

1. Chang MN, Therneau TM, Wieand HS, Cha SS. Designs for group sequential phase II clinical trials. *Biometrics*. 1987; 43(4):865–874. [PubMed: 3427171]
2. Chen TT. Optimal three-stage designs for phase II cancer clinical trials. *Statistics in Medicine*. 1997; 16(23):2701–2711. [PubMed: 9421870]
3. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. *Statistics in Medicine*. 1994; 13(17):1727–1736. [PubMed: 7997706]
4. Jung SH, Carey M, Kim KM. Graphical search for two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 2001; 22:367–372. [PubMed: 11514038]

5. Shuster J. Optimal two-stage designs for single arm phase II cancer trials. *Journal of Biopharmaceutical Statistics*. 2002; 12(1):39–51. [PubMed: 12146719]
6. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
7. Therneau TM, Wieand HS, Chang MN. Optimal designs for a grouped sequential binomial trial. *Biometrics*. 1990; 46:771–783.
8. London MB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics In Medicine*. 2005; 24:2597–2611. [PubMed: 16118809]
9. Jung SH, Chang MN, Kang SJ. Phase II Clinical Trials with Heterogeneous Patient Populations. *Journal of Biopharmaceutical Statistics*. 2011 to appear.
10. Sposto R, Gaynon PS. An adjustment for patient heterogeneity in the design of two-stage phase II trials. *Statistics in Medicine*. 2009; 28:2566–2579. [PubMed: 19521973]
11. Thall PF, Wathen JK, Bekele BN, Champlin RE, Baker LH, Benjamin RS. Hierarchical Bayesian approaches to phase II trials in disease with multiple subtypes. *Statistics in Medicine*. 2003; 22(5): 763–780. [PubMed: 12587104]
12. Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. *Statistics in Medicine*. 2008; 27:2802–2815. [PubMed: 17948869]
13. Chang MN, Jung SH, Wu SS. Two-stage designs with additional futility tests for phase II clinical trials with heterogeneous patient population. *Sequential Analysis*. 2011; 30:338–349.
14. Cox, DR.; Hinkley, DV. *Theoretical Statistics*. Chapman and Hall; London: 1974.
15. Chang MN, Hwang IK, Shih WJ. Group sequential designs using both Type I and Type II error probability spending functions. *Communications in Statistics, Theory and Methods*. 1998; 27(6): 1323–1339.
16. London WB, Frantz CN, Campbell LA, Seeger RC, Brumback BA, Cohn SL, Matthay KK. Phase II Randomized Comparison of Topotecan plus Cyclophosphamide Versus Topotecan Alone in Children with Recurrent or Refractory Neuroblastoma: A Children's Oncology Group Study. *Journal of Clinical Oncology*. 2010; 28(24):3808–3815. [PubMed: 20660830]

Table1

Examples of two-stage design¹

P_{10}	P_{20}	P_{30}	M_f	n_{11}	n_{21}	n_{31}	Propose tests						Tests by London and Chang					
							a	b	c	a	$1-\beta$ under H_a^2	$1-\beta$ under H_a^{*3}	a	b	c	a	$1-\beta$ under H_a^2	$1-\beta$ under H_a^{*3}
0.5	0.3	0.1	20	2	4	14	4.394	8.287	11.834	0.048	0.912	0.964	4	7	11	0.039	0.895	0.953
				4	6	10	5.241	9.134	13.371	0.049	0.892	0.937	5	9	14	0.031	0.851	0.898
				6	7	7	5.931	9.479	14.907	0.049	0.873	0.895	6	10	17	0.041	0.769	0.782
				14	4	2	7.625	11.518	18.799	0.047	0.838	0.623	9	13	22	0.033	0.786	0.522
				10	6	4	6.778	10.671	16.946	0.050	0.859	0.794	8	12	19	0.039	0.828	0.723
				7	7	6	5.931	9.823	15.410	0.049	0.871	0.877	7	11	17	0.034	0.827	0.810
			18	2	4	12	3.891	7.942	10.987	0.050	0.891	0.950	4	7	11	0.024	0.813	0.899
				4	5	9	4.739	8.287	12.523	0.047	0.862	0.910	5	8	13	0.032	0.813	0.856
				6	6	6	5.083	8.976	14.060	0.045	0.839	0.855	6	10	15	0.036	0.806	0.797
				12	4	2	6.778	10.326	16.946	0.045	0.795	0.617	8	12	19	0.050	0.809	0.591
				9	5	4	5.931	9.823	15.410	0.045	0.820	0.751	7	11	17	0.043*	0.807	0.701
			16	2	3	11	3.389	7.095	10.139	0.046	0.853	0.923	3	6	10	0.026	0.776	0.866
				3	5	8	3.891	7.784	11.173	0.048	0.837	0.899	4	8	12	0.029	0.713	0.789
				5	5	6	4.236	8.129	12.365	0.050	0.820	0.847	5	8	15	0.049	0.570	0.585
				11	3	2	5.931	9.479	15.252	0.049	0.769	0.587	7	11	18	0.026	0.669	0.423
				8	5	3	5.083	8.976	14.060	0.044	0.771	0.702	6	10	16	0.032	0.717	0.605
				6	5	5	4.739	8.473	12.868	0.047	0.801	0.802	5	9	14	0.037	0.764	0.733
0.4	0.2	0.1	20	2	4	14	4.122	8.003	10.945	0.049	0.911	0.962	4	7	10	0.026	0.871	0.939
				4	6	10	4.491	8.275	12.028	0.048	0.894	0.941	4	8	12	0.026	0.850	0.898
				6	7	7	4.763	8.687	12.839	0.050	0.885	0.911	5	9	13	0.042	0.871	0.877
				14	4	2	5.846	9.800	15.476	0.048	0.839	0.662	7	11	18	0.039	0.811	0.567
				10	6	4	5.404	9.158	14.194	0.049	0.860	0.813	6	10	16	0.049*	0.812	0.707
				7	7	6	4.933	8.814	13.281	0.048	0.875	0.889	5	9	14	0.035	0.846	0.832
			18	2	4	12	3.510	7.362	10.334	0.047	0.888	0.950	3	6	9	0.042	0.880	0.941

Propose tests							Tests by London and Chang												
							a	b	c	a	1-β under Ha ²	1-β under Ha ^{*3}	a	b	c	a	1-β under Ha ²	1-β under Ha ^{*3}	
P ₁₀							n ₃₁	a	b	c	a	1-β under Ha ²	1-β under Ha ^{*3}	a	b	c	a	1-β under Ha ²	1-β under Ha ^{*3}
							9	3.952	7.707	11.145	0.050	0.874	0.922	4	7	11	0.032	0.828	0.870
							6	4.224	8.076	12.028	0.048	0.853	0.872	5	8	12	0.050	0.847	0.840
							2	5.035	8.989	14.024	0.047	0.806	0.647	6	10	16	0.041	0.786	0.570
							4	4.763	8.547	12.941	0.050	0.836	0.788	5	9	15*	0.046	0.731*	0.619
							11	2.971	6.823	9.426	0.049	0.864	0.932	3	6	8	0.044	0.855	0.918
							8	3.141	7.095	10.037	0.049	0.847	0.909	3	7	10	0.023	0.759	0.827
							6	3.510	7.367	10.775	0.050	0.827	0.858	4	7	11	0.037	0.782	0.786
							2	4.224	8.178	12.703	0.050	0.775	0.611	5	9	15	0.031	0.694	0.459
							3	4.054	7.809	11.863	0.049	0.794	0.742	5	9	13	0.036	0.750	0.645
							5	3.583	7.537	11.120	0.050	0.819	0.829	4	8	12	0.040*	0.734	0.705
	0.3	0.2	0.1	20	2	4	14	4.025	7.731	10.775	0.048	0.910	0.962	3	7	9	0.042	0.915	0.965
					4	6	10	4.158	8.051	11.387	0.050	0.899	0.944	4	7	11	0.032	0.863	0.909
					6	7	7	4.503	8.185	11.975	0.049	0.884	0.910	4	8	12	0.035	0.861	0.872
					14	4	2	5.083	8.741	13.480	0.049	0.847	0.670	6	10	15	0.048	0.815	0.578
					10	6	4	4.637	8.529	12.766	0.050	0.867	0.823	5	9	14	0.028	0.809	0.707
				7	7	6	4.503	8.318	12.210	0.049	0.880	0.893	5	8	12	0.048	0.874	0.865	
			18	2	4	12	3.547	7.228	10.037	0.047	0.885	0.948	3	6	9	0.028	0.848	0.923	
				4	5	9	3.656	7.471	10.540	0.049	0.875	0.921	3	7	10	0.033	0.845	0.886	
				6	6	6	3.789	7.682	11.127	0.050	0.859	0.879	4	8	11	0.039	0.836	0.833	
				12	4	2	4.370	8.051	12.263	0.049	0.818	0.668	5	9	14	0.041	0.740	0.513	
				9	5	4	4.025	7.918	11.707	0.049	0.840	0.792	4	8	12	0.048	0.839	0.751	
			16	2	3	11	2.699	6.726	9.190	0.047	0.863	0.932	3	6	8	0.029	0.815	0.893	
				3	5	8	3.044	6.859	9.559	0.049	0.847	0.909	3	6	9	0.034	0.803	0.863	
				5	5	6	3.178	7.070	10.013	0.049	0.831	0.864	3	7	10	0.031	0.779	0.787	
				11	3	2	3.522	7.415	11.149	0.050	0.786	0.620	4	8	13	0.044	0.663	0.426	
				8	5	3	3.522	7.204	10.727	0.048	0.798	0.746	4	8	12	0.032	0.680	0.570	
				6	5	5	3.178	7.179	10.248	0.050	0.824	0.834	3	7	10	0.048	0.819	0.797	

P_{10}	P_{20}	P_{30}	M_j	n_{11}	n_{21}	n_{31}	Propose tests						Tests by London and Chang					
							a	b	c	a	$1-\beta$ under H_a^2	$1-\beta$ under H_a^{*3}	a	b	c	a	$1-\beta$ under H_a^2	$1-\beta$ under H_a^{*3}
0.2	0.15	0.1	20	2	4	14	3.815	7.497	10.196	0.048	0.915	0.966	3	6	8	0.043	0.923	0.968
				4	6	10	3.815	7.631	10.609	0.050	0.908	0.948	3	6	9	0.040	0.906	0.942
				6	7	7	4.058	7.874	10.951	0.048	0.901	0.920	4	7	10	0.026	0.859	0.873
				14	4	2	4.058	8.117	11.663	0.049	0.884	0.724	4	8	11	0.044	0.882	0.681
				10	6	4	4.192	7.874	11.329	0.049	0.889	0.844	4	7	11	0.028	0.845	0.757
				7	7	6	4.192	7.874	11.051	0.050	0.898	0.908	4	7	10	0.033	0.873	0.866
			18	2	4	12	3.446	7.262	9.593	0.049	0.898	0.954	3	6	8	0.025	0.856	0.928
				4	5	9	3.446	7.128	9.827	0.050	0.888	0.930	3	6	8	0.046	0.892	0.924
				6	6	6	3.346	7.262	10.105	0.049	0.877	0.896	3	6	9	0.038	0.862	0.862
				12	4	2	3.923	7.370	10.682	0.049	0.855	0.705	4	7	10	0.043	0.845	0.658
				9	5	4	3.446	7.405	10.447	0.049	0.866	0.818	3	7	10	0.026	0.816	0.723
			16	2	3	11	2.465	6.516	8.747	0.049	0.877	0.941	2	5	7	0.036	0.857	0.923
				3	5	8	2.465	6.659	9.081	0.050	0.867	0.920	2	6	8	0.022	0.794	0.859
				5	5	6	2.942	6.758	9.224	0.047	0.849	0.874	3	6	8	0.035	0.823	0.832
				11	3	2	2.942	7.001	9.809	0.050	0.831	0.660	3	7	9	0.045	0.824	0.620
				8	5	3	3.077	6.758	9.601	0.046	0.828	0.780	3	6	9	0.032	0.790	0.697
				6	5	5	3.077	6.758	9.324	0.050	0.849	0.851	3	6	8	0.045	0.841	0.823

$I_k = 3$; $\Delta_1 = \Delta_2 = \Delta_3 = 0.2$; $n_1 = n_2$, $i=1, 2, 3$; $H_0: \mathbf{p} = \mathbf{p}_0 = (p_{10}, p_{20}, p_{30})$ vs. $H_a: \mathbf{p} = \mathbf{p}_1 = (p_{10} + \Delta_1, p_{20} + \Delta_2, p_{30} + \Delta_3)$. Design parameters are defined in the first paragraph of Section 3.

² The power under H_a : $\mathbf{p} = \mathbf{p}_1 = (p_{10} + \Delta_1, p_{20} + \Delta_2, p_{30} + \Delta_3)$.

³ The power under H_a^* : $\mathbf{p} = \mathbf{p}_2 = (p_{10} + \Delta - 0.1, p_{20} + \Delta_2 + 0.05, p_{30} + \Delta_3 + 0.05)$. The weights w_j 's for the proposed tests are computed by (1) using H_0 vs. H_a .

* These numbers are different from that in Table IV in London and Chang [8]. We guess that there are typos in London and Chang [8].