Published in final edited form as:

Contemp Clin Trials. 2015 May ; 42: 9–17. doi:10.1016/j.cct.2015.02.007.

Statistical Issues for Design and Analysis of Single-Arm Multi-Stage Phase II Cancer Clinical Trials

Sin-Ho Jung

Biostatistics and Clinical Epidemiology Center, Samsung Medical Center Department of Biostatistics and Bioinformatics, Duke University

Summary

Phase II trials have been very widely conducted and published every year for cancer clinical research. In spite of the fast progress in design and analysis methods, single-arm two-stage design is still the most popular for phase II cancer clinical trials. Because of their small sample sizes, statistical methods based on large sample approximation are not appropriate for design and analysis of phase II trials. As a prospective clinical research, the analysis method of a phase II trial is predetermined at the design stage and it is analyzed during and at the end of the trial as planned by the design. The analysis method of a trial should be matched with the design method. For two-stage single arm phase II trials, Simon's method has been the standards for choosing an optimal design, but the resulting data have been analyzed and published ignoring the two-stage design aspect with small sample sizes. In this article, we review analysis methods that exactly get along with the exact two-stage design method. We also discuss some statistical methods to improve the existing design and analysis methods for single-arm two-stage phase II trials.

Keywords

Accrual overrun; Admissible design; Confidence interval; P-value; Two-stage design; Uniformly minimum unbiased estimator

1 Introduction

Cancer clinical trials are to investigate the efficacy and toxicity of experimental cancer therapies. If an appropriate dose level of an experimental drug is determined from a phase I trial, the drug's anticancer activity is assessed through phase II clinical trials. Phase II clinical trials are to screen out inefficacious experimental therapies before they proceed to further investigation through large scale phase III trials. In order to expedite this process, phase II trials traditionally use a single-arm design to treat patients by experimental

© 2015 Published by Elsevier Inc.

Corresponding Author: Sin-Ho Jung, Ph.D., Department of Biostatistics and Bioinformatics, Duke University, 2424 Erwin Road, 11070 Hock Plaza, Suite 1102, DUMC Box 2721, Durham, NC 27710, U.S.A., sinho.jung@duke.edu, phone: 919-668-8658, fax: 919-668-5888.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

therapies only. So, the efficacy of an experimental therapy is compared with that of a standard therapy that is called a historical control. The most popular primary endpoint of phase II cancer clinical trials is tumor response which is measured by the change in tumor size before and during treatment. For a solid tumor, if the size, defined as sum of the largest diameter of the target tumor decreases by at least 30% compared to that at the baseline, we call it a partial response by RECIST 1.1 [1]. If the target tumor disappears during treatment, then we call it a complete response. Overall response is defined as partial or complete response.

Phase II trials generally require shorter study periods than phase III trials. Consequently, phase II trials have small sample sizes, so that exact statistical methods are preferable to the asymptotic methods for their design and analysis. Various exact methods have been published for phase II trials with binary outcomes such as tumor response. For ethical reasons, two-stage designs are commonly used for phase II cancer clinical trials. A typical single-arm two-stage trial with a futility (a_1) and a superiority (b_1) stopping values and a stage 2 rejection value (a) are conducted as follows.

- Stage 1: Treat n_1 patients and count the number of responders X_1 .
 - If $X_1 \leq a_1$, then reject the experimental therapy and stop the trial.
 - If $X_1 \geq b_1$, then accept the experimental therapy and stop the trial.
 - If $a_1 < X_1 < b_1$, then proceed to the second stage.
- Stage 2: Treat an additional n_2 patients and count the number of responders X_2 .
 - If $X_1 + X_2 \geq a$, then reject the experimental therapy.
 - Otherwise, accept the experimental therapy for further investigation.

We usually do not stop the trial early for superiority since there is no ethical issue to continue treating patients with an efficacious therapy and we want to collect as much data as possible to use when designing a subsequent phase III trial. In this case, we choose $b_1 = n_1 + 1$.

The design and analysis of phase II trials require exact statistical methods accounting for two-stage design and small sample sizes. The most popular for phase II cancer clinical trials has been a two-stage design with a futility stopping only based on Simon's [2] minimax or optimality criterion [3-5]. But most publications reporting the results of phase II trials fail to appropriately address these issues. When the above two-stage phase II trial is completed, we should be able to accept or reject the experimental therapy based on the critical values and sample sizes, (a_1, b_1, a, n_1, n_2). In a usual clinical trial, we would recruit slightly more patients than required to make up for possible attrition due to ineligibility and dropout. As a result, the observed sample size tends to be different from that specified by the design. In a typical two-stage phase II trial, the sample size at the stopping stage is different from (usually larger, but rarely smaller, than) the planned one. In this case, the critical values at the stopping stage become meaningless. This may be the major reason why investigators do not make clear conclusions from their phase II trials. Addressing this issue, we propose to extend the two-stage testing rule by calculating a p-value or a confidence interval of the true

response rate (RR). These methods exactly coincide with the two-stage testing rule if the observed sample sizes are identical to the prespecified (n_1, n_2) . In this article, we discuss further design and analysis methods for single-arm phase II trials with tumor response as the primary endpoint. We further extend the Simon's criteria to search for an alternative two-stage designs.

2 Optimal Two-Stage Designs

In this section, we focus on the popular two-stage designs with a futility interim test only i.e. $b_1 = n_1 + 1$. These designs are defined by the number of patients to be treated during stages 1 and 2, n_1 and n_2 , and rejection values a_1 and a ($a_1 < a$), so that we specify them by $(a_1/n_1, a/n)$, where $n = n_1 + n_2$ is called the maximal sample size. The values of $(a_1/n_1, a/n)$ are determined based on some prespecified design parameters as described below. Let p_0 denote the maximum unacceptable probability of response which is usually chosen by the RR of a historical control, and p_1 the minimum acceptable probability of response with $p_0 < p_1$. For the true RR p of the experimental therapy, we want to test $H_0 : p \leq p_0$ against $H_1 : p > p_0$. In this statistical testing, rejecting H_0 means accepting the experimental therapy. Given (p_0, p_1) , we can calculate the type I error probability α and power $1 - \beta$ of a two-stage design $(a_1/n_1, a/n)$ based on the fact that the number of responders from the two stages, X_1 and X_2 , are independent binomial random variables.

Suppose that we want to find a two-stage design with a type I error probability no larger than α^* and a power no smaller than $1 - \beta^*$ for given (p_0, p_1) values. We call $(p_0, p_1, \alpha^*, \beta^*)$ design parameters. Given (p_0, p_1) , there are infinitely many 2-stage designs satisfying the $(\alpha^*, 1 - \beta^*)$ -constraint. Noting that the sample size of a two-stage trial is a random variable taking n_1 or n , we can calculate the expected sample size of the trial given a true RR of the experimental therapy. Simon [2] proposes two criteria to select a good two-stage design among these candidate designs. The minimax design minimizes the maximal sample size, n , among the designs satisfying the $(\alpha^*, 1 - \beta^*)$ -constraint. On the other hand, the so-called "optimal" design minimizes the expected sample size under $H_0 : p = p_0$, denoted as EN.

While the minimax and the optimal designs have both been widely used for long, other two-stage designs have been largely ignored in the past. Oftentimes, the two Simon's criteria result in very different two-stage designs, i.e. the minimax design may have an excessively large EN as compared to the optimal design and the optimal design may have an excessively large maximal sample size n as compared to the minimax design. This results from the discrete nature of the exact binomial method.

Example 1: For the design parameters $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.1, 0.3, 0.05, 0.85)$, the minimax design is given by $(a_1/n_1, a/n) = (2/18, 5/27)$ and the optimal design by $(1/11, 6/35)$. The maximal sample size n for the minimax design is less than that for the optimal design by 8. However, the expected sample size EN under H_0 for the optimal design is 18.3 which is only slightly smaller than $EN = 20.4$ for the minimax design. We may consider choosing the minimax design since, compared to the optimal design, it largely saves the maximal sample size while sacrificing the expected sample size by only about 2.

In this case, however, there exists a practical compromise without changing the statistical operating characteristics appreciably.

Example 1 (revisited): For the same design parameters $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.1, 0.3, 0.05, 0.85)$, the design given by $(a_1/n_1, a/n) = (1/13, 5/28)$ requires only one more patient in the maximal sample size n than the minimax design but its expected sample size EN under H_0 is very comparable to that of the optimal design (18.7 vs 18.3).

This design is a good compromise between the minimax design and the optimal design [6]. There can be multiple compromising designs. Jung et al. [7] show that they are admissible designs in terms of the loss function combining maximal sample size and the expected sample size under H_0 . Simon's minimax and optimal designs belong to the family of admissible designs too. An interactive computer program to find admissible designs is available in www.dukecancerinstitute.org/research/shared-resources/biostatistics/clinical-trial-design-systems/.

Figure 1 is a snapshot of this program for Example 1. It displays the plot of EN against n for various designs with $n \leq 37$. Simon's minimax design given by $(a_1/n_1, a/n) = (2/18, 5/27)$ comes the leftmost and the optimal design by $(1/11, 6/35)$ comes at the very bottom. The program provides specification of a design $(a_1/n_1, a/n)$ along with $(\alpha, 1 - \beta)$, EN, and the probabilities of early termination for $p = p_0$ and p_1 when the circle representing a candidate design, actually (n, EN) , is clicked with a pointer.

In this section, we have considered two-stage designs with a futility stopping value only. Chang et al. [8] propose optimal multistage designs with both futility and superiority stopping boundaries by minimizing the average of expected sample sizes under H_0 and H_1 . Mander et al. [9] extends the admissible design concept to multistage designs with both futility and superiority stopping boundaries by using a loss function combining maximal sample size, expected sample size under H_0 and expected sample size under H_1 . Englert and Kieser [10] propose optimal adaptive two-stage designs whose stage 2 sample sizes are determined based on the number of responders from stage 1. They do not consider early stopping. Accounting for uncertainty of the true treatment effect, Kieser [11] proposes to specify an interval of possible true response rate in designing a two-stage design.

3 Inference on Response Rate

3.1 Estimation of Response Rate

In a two-stage phase II trial, let M ($=1$ or 2) denote the stopping stage, and S the cumulative number of responders by the stopping stage, i.e. $S = X_1$ if $M = 1$ and $S = X_1 + X_2$ if $M = 2$. The most popular estimator of RR p for $(M, S) = (m, s)$ is the sample proportion, i.e.

$$\hat{p} = \begin{cases} s/n_1 & \text{if } m=1 \\ s/(n_1+n_2) & \text{if } m=2 \end{cases}$$

By not reflecting the two-stage design aspect of the trial, this estimator, called the maximum likelihood estimator (MLE), is always negatively biased for standard two-stage trials with futility stopping only [12]. As a result, if p_0 of a future trial is chosen by the sample proportion of a historical control taken from a previous two-stage phase II trial, then it will underestimate the true RR of the historical control and result in a higher chance of false positivity for the future trial.

For $(M, S) = (m, s)$, the uniformly minimum variance unbiased estimator (UMVUE) of p for two-stage phase II trials is given by

$$\tilde{p} = \begin{cases} s/n_1 & \text{if } m=1 \\ \frac{\sum_{x_1=(a_1+1) \vee (s-n_2)}^{s \wedge (b_1-1)} \binom{n_1-1}{x_1-1} \binom{n_2}{s-x_1}}{\sum_{x_1=(a_1+1) \vee (s-n_2)}^{s \wedge (b_1-1)} \binom{n_1}{x_1} \binom{n_2}{s-x_1}} & \text{if } m=2 \end{cases} \quad (1)$$

where $a \wedge b = \min(a, b)$, $a \vee b = \max(a, b)$, $\binom{m}{k} = m! / \{k!(m-k)!\}$ and $x! = x \times (x-1) \times \dots \times 2 \times 1$ [12]. Note that the UMVUE and the MLE are identical if the trial stops after stage 1, i.e. $m = 1$.

For a true RR of p , the probability mass function (PMF) of (M, S) , $f(m, s|p) = \Pr(M = m, S = s)$, is given as [12]

$$f(m, s|p) = \begin{cases} p^s(1-p)^{n_1-s} \binom{n_1}{s} & \text{if } m=1, 0 \leq s \leq a_1 \text{ or } b_1 \leq s \leq n_1 \\ p^s(1-p)^{n_1+n_2-s} \sum_{x_1=a_1+1}^{(b_1-1) \wedge s} \binom{n_1}{x_1} \binom{n_2}{s-x_1} & \text{if } m=2, a_1+1 \leq s \leq b_1-1+n_2 \end{cases} \quad (2)$$

Since the UMVUE of p is a function of (M, S) , its PMF is derived from $f(m, s|p)$.

Example 2: Under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.2, 0.4, 0.05, 0.8)$, the Simon's optimal two-stage design with a futility stopping value only is given as $(a_1/n_1, a/n) = (3/13, 12/43)$. Table 1 gives the UMVUE and the MLE for observations (M, S) for the two-stage design and their PMF for various RR values. When $m = 1$, two estimates are exactly the same as noted earlier. When $m = 2$, the MLE is much smaller than UMVUE for small s values. The UMVUE is shown to have a comparable variance as compared to the MLE overall [12].

In order to account for ineligible or unevaluable patients, we often accrue slightly more (or possibly fewer) patients than the planned sample size at the stopping stage, especially in multicenter trials [13,14]. But, this does not become any issue in the calculation of the UMVUE, since it does not require specification of the critical values (a_m, b_m) at the stopping stage m , where $a_2 = a$ and $b_2 = a + 1$. If the study is terminated after stage 1 (i.e., $m = 1$), then the UMVUE is calculated by regarding the observed number of stage 1 patients as n_1 . If the study is proceeded to the second stage (i.e., $m = 2$), then the UMVUE in (1) depends

only on the rejection values for stage 1 (a_1, b_1) and the number of patients for the two stages (n_1, n_2) . When $m = 2$, the interim test is always conducted using (a_1, b_1) when exactly n_1 patients have response data as planned. However, the number of stage 2 patients may be slightly different from n_2 . In this case, we calculate the UMVUE by regarding the observed number of patients for stage 2 as n_2 with (n_1, a_1, b_1) fixed by the design.

Example 2 (revisited): Suppose that a phase II trial was designed for the Simon's optimal two-stage design with a futility stopping value only is given as $(a_1/n_1, a/n) = (3/13, 12/43)$ under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.2, 0.4, 0.05, 0.8)$, but it was completed with $s = 7$ responders from a total of 45 patients after stage 2. Then, by using $(a_1, n_1) = (3, 13)$ and conditioning n_2 at $32 (= 45 - 13)$, we have $(a_1 + 1) \vee (s - n_2) = (3 + 1) \vee (7 - 32) = 4$ and $s \wedge (b_1 - 1) = 7 \wedge (14 - 1) = 13$. Hence, from (1), the UMVUE is calculated by

$$\tilde{p} = \frac{\sum_{x_1=4}^7 \binom{12}{x_1-1} \binom{32}{7-x_1}}{\sum_{x_1=4}^7 \binom{13}{x_1} \binom{32}{7-x_1}} = \frac{1,362,988}{4,241,380} = 0.321$$

while the MLE is only $\hat{p} = 7/45 = 0.156$.

Bowden and Wason [15] investigate various estimators of response rate including the UMVUE and compare their performance.

3.2 Confidence Interval

Jung and Kim [12] proves that the ordering of the sample space for (M, S) in terms of the magnitude of the UMVUE is identical to that by Jennison and Turnbull [16], see also Armitage [17]; Tsiatis, Rosner and Mehta [18]. In other words, we have

$$\begin{aligned} \tilde{p}(1, 0) &< \tilde{p}(1, 1) < \cdots < \tilde{p}(1, a_1) \\ &< \tilde{p}(2, a_1+1) < \cdots < \tilde{p}(2, b_1-1+n_2) \\ &< \tilde{p}(1, b_1) < \cdots < \tilde{p}(1, n_1) \end{aligned} \quad (3)$$

where $\tilde{p}(m, s)$ denotes the UMVUE for $(M, S) = (m, s)$. Hence, the confidence intervals calculated according to Clopper and Pearson [19] and the stochastic ordering based on the magnitude of the UMVUE are identical to those by Jennison and Turnbull [16].

With the UMVUE $\tilde{p}(m, s)$, an exact $100(1 - \alpha)\%$ equal tail confidence interval (p_L, p_U) for p is given by

$$\Pr(\tilde{p}(M, S) \geq \tilde{p}(m, s) | p = p_L) = \alpha/2$$

and

$$\Pr(\tilde{p}(M, S) \leq \tilde{p}(m, s) | p = p_U) = \alpha/2,$$

where the probabilities are calculated using the PMF of (M, S) as defined in (2). Confidence limits p_L and p_U can be obtained by solving the equations using a numerical method such as the bisection method which can be described to solve an equation $g(x) = 0$ as:

- A. Find an interval $[x_1, x_2]$ such that $g(x_1)g(x_2) < 0$.
- B. For $x_3 = (x_1 + x_2)/2$,
 - a. If $g(x_1)g(x_3) < 0$, then $x_3 \rightarrow x_2$.
 - b. Otherwise, $x_3 \rightarrow x_1$.
- C. Continue (B) until $|x_1 - x_2|$ is small enough, and choose x_3 as the solution to the equation.

Example 3: Suppose that we observed $(m, s) = (2, 7)$ from a two-stage study with lower boundaries only, $(n_1, n_2, a_1, a) = (13, 30, 3, 12)$. In this case, we have $b_1 = n_1 + 1 = 14$, $(a_1 + 1) \vee (s - n_2) = (3 + 1) \vee (7 - 30) = 4$, and $s \wedge (b_1 - 1) = 7 \wedge (14 - 1) = 7$. From (1), the UMVUE is

$$\tilde{p}(2, 7) = \frac{\sum_{x_1=4}^7 \binom{13-1}{x_1-1} \binom{30}{7-x_1}}{\sum_{x_1=4}^7 \binom{13}{x_1} \binom{30}{7-x_1}} = .322.$$

Using (2), we have

$$\begin{aligned} \Pr(\tilde{p}(M, S) \geq .322 | p = .103) &= .025, \\ \Pr(\tilde{p}(M, S) \leq .322 | p = .538) &= .025 \end{aligned}$$

so that a 95% confidence interval on p is given as $(.103, .538)$, which is the same as the one according to Jennison and Turnbull [16]. In contrast, a naive exact 95% confidence interval by Clopper and Pearson [19] ignoring the two-stage aspect of the study design is given as $(.068, .307)$. Note that the latter is narrower than the former by ignoring the group sequential feature of the study. Furthermore, the former is slightly shifted to the right from the latter to reflect the fact that the study has been continued to stage 2 after observing more responders than $a_1 = 3$ in stage 1. The popularly used asymptotic confidence interval

$\hat{p} \pm 1.96 \sqrt{\hat{p}(1 - \hat{p})/n} = (.052, .0273)$ with 95% significance level is even narrower and further shifted to the left than the naive exact confidence interval.

The Jennison-Turnbull confidence interval based on the stochastic ordering (3) has a desirable property: given (p_0, α^*) , the testing result based on the two-stage design $(a_1/n_1, a/n)$ exactly coincides with that based on the one-sided confidence interval with significance

level $100(1 - \alpha^*)\%$, i.e. we reject $H_0 : p = p_0$ by the two-stage design if and only if the lower confidence limit is larger than p_0 . This property is very valuable especially when the observed sample size for the stopping stage is different from the sample size specified by the study design. In this case, the two-stage design does not provide us a proper testing rule any more since the rejection values, (a_1, b_1) or a , are meaningful only when the number of patients are identical to the prespecified sample size. But we can calculate a Jennison-Turnbull confidence interval of RR by treating the observed sample size at the stopping stage like the planned sample size. This is possible since (2) and (3) do not require specification of the rejection values at the stopping stage. Various confidence interval methods have been proposed for multistage designs [20,21], but no other confidence intervals have this property.

Example 4: Kim et al. [22] published a phase II trial of concurrent chemoradiotherapy for newly diagnosed patients with limited-stage extranodal natural killer/T-cell lymphoma, nasal type. The primary endpoint is complete response (CR) rate. The study had the Simon's optimal two-stage design $(a_1/n_1, a/n) = (4/6, 22/27)$ under $(p_0, p_1, \alpha^*, 1 - \beta^*) = (0.7, 0.9, 0.05, 0.8)$. The trial proceeded to the second stage to treat a total of 30 eligible and evaluable patients of which 27 achieved CR. Since the observed maximal sample size is different from the planned one, the rejection value $a = 22$ was of no use to determine the positivity of the study. The authors concluded the study to be positive without statistical ground. Noting this, Shimada and Suzuki [23] claimed that this trial was a negative study by calculating an asymptotic confidence interval with 2-sided 95% significance level and showing that it covers $p_0 = 0.7$. This claim is misleading since their interval (i) has a two-sided $100(1 - \alpha^*) = 95\%$ confidence level while the two-stage design has 1-sided $\alpha^* = 5\%$, (ii) ignores the two stage feature of the trial, and (iii) is using a large-sample approximation. Kim et al [24] defended their conclusion by showing that the 1-sided 95% Jennison-Turnbull confidence interval $(0.762, 1]$ is above $p_0 = 0.7$.

4 P-Value Calculation

Through a phase II trial, we intend to conduct a statistical testing to reject or accept its experimental therapy. If we reject or fail to reject the null hypothesis, we should be able to provide a p-value as a measure of how strong evidence the decision is based on against the null hypothesis. However, publications from phase II trials rarely report p-values to support their conclusions. In most of the publications, it is not even clear if the investigators conclude to accept their therapies or not.

Using the stochastic ordering of UMVUE (3), Jung et al. [25] propose a p-value method for 2-stage phase II clinical trials. Noting that a p-value is defined as the probability to observe an extreme test statistic value toward the direction of H_1 when H_0 is true, they propose to calculate the probability to observe a UMVUE value larger than that obtained from the study under H_0 . Let $p^{\sim}$ denote the UMVUE of the RR observed from a two-stage phase II trial specified by (a_1, b_1, a, n_1, n_2) . Given $(M, S) = (m, s)$, the p-value $= Pr\{p^{\sim}(M, S) \geq p^{\sim}(m, s) | p_0\}$ based on UMVUE can be calculated as

$$p - \text{value} = \begin{cases} \sum_{j=s}^{n_1} f(1, j|p_0) & \text{if } m=1, s \geq b_1 \\ 1 - \sum_{j=0}^{s-1} f(1, j|p_0) & \text{if } m=1, s \leq a_1 \\ \sum_{j=b_1}^{n_1} f(1, j|p_0) + \sum_{j=s}^{b_1-1+n_2} f(2, j|p_0) & \text{if } m=2 \end{cases} \quad (4)$$

Jung et al. [25] show that, if the observed sample size at each stage is identical to that specified by the design, the decision by the two-stage design exactly matches with that by the p-value compared with the α^* value. As discussed in the previous section, the confidence interval by Jennison and Turnbull [16] has this property too.

Example 5: For $H_0 : p_0 = .4$ vs. $H_1 : p_1 = .6$, the two-stage design $(n_1, n_2, a_1, b_1, a) = (34, 5, 17, 35, 20)$ is the Simon's minimax design among those with $(\alpha^*, 1 - \beta^*) = (0.05, 0.8)$ and a futility stopping boundary only. Table 2 displays the p-values and the PMF at $p_0 = 0.4$ for each sample point in the descending order for UMVUE. For example, for $(m, s) = (2, 19)$, we calculate the UMVUE by

$$\tilde{p} = \frac{\sum_{x_1=(17+1)\vee(19-5)}^{19\wedge(35-1)} \binom{34-1}{x_1-1} \binom{5}{19-x_1}}{\sum_{x_1=(17+1)\vee(19-5)}^{19\wedge(35-1)} \binom{34}{x_1} \binom{5}{19-x_1}} = \frac{\binom{33}{17} \binom{5}{1} + \binom{33}{18} \binom{5}{0}}{\binom{34}{18} \binom{5}{1} + \binom{34}{19} \binom{5}{0}} = 0.5337.$$

Hence, when $(m, s) = (2, 19)$ is observed, we calculate the p-value based on UMVUE as

$$\begin{aligned} \Pr(\tilde{p}(M, S) \geq .5337|p_0) \\ &= \sum_{j=19}^{39} f(2, j|p_0) \\ &= .0129 + .0219 + .0217 + .0145 + .0075 + .0033 + .0013 + .0005 + .0002 + .0000 + \cdots + .0000 = .0838 \end{aligned}$$

Since p-value $> \alpha^* (= 0.05)$, we fail to reject H_0 . Since $s = 19$ after stage 2 ($= m$) is smaller than $a = 20$, we fail to reject H_0 by the two stage testing rule too. Table 2 also reports the p-values based on the ordering by MLE values and the p-values calculated by ignoring the fact that the two-stage design aspect, denoted as 'naive'. Note that the p-values based on UMVUE and MLE orderings are different for large s values with $m = 1$ and for small s values with $m = 2$. This occurs because of the difference between the UMVUE-ordering and the MLE-ordering for the small s values with $m = 2$. The naive p-values are quite different from the p-values based on the UMVUE-ordering too.

The strength of the above p-value method is that it can be extended to the cases where the observed sample size at the stopping stage is different from that specified by the design. This becomes possible because the PMF $f(m, s|p)$ and the UMVUE depend on the stopping boundaries only up to stage $m - 1$, i.e. $\{(a_k, b_k), 1 < k \leq m - 1\}$, where $a_2 = b_2 - 1$ denotes a , the rejection value at the second stage. Suppose that a study stopped after stage $m = 1$ with S

= s responders from n_1 patients that may be different from the stage 1 sample size specified by the design. Then, (4) can be modified to calculate a p-value

$$\text{p-value} = \begin{cases} \sum_{j=s}^{n_1} p^s (1-p)^{n_1-s} \binom{n_1}{s} & m=1 \text{ and the study is stopped for superiority} \\ 1 - \sum_{j=0}^{s-1} p^s (1-p)^{n_1-s} \binom{n_1}{s} & m=1 \text{ and the study is stopped for futility} \end{cases}$$

Note that the p-value with $m = 1$ is calculated as if the study had a single-stage design with sample size n_1 , so that it does not require specification of stopping values (a_1, b_1).

On the other hand, suppose that the study proceeded to the second stage, i.e. $m = 2$. In this case, the stage 1 decision will be made based on (a_1, b_1, n_1) as specified by the design, but the sample size for the second stage may be slightly different from n_2 that is specified by the design. In this case, based on observed values of (n_2, s) , we calculate

$$\text{p-value} = \sum_{j=b_1}^{n_1} p^s (1-p)^{n_1-s} \binom{n_1}{s} + \sum_{j=s}^{b_1-1+n_2} p^s (1-p)^{n_1+n_2-s} \sum_{x_1=a_1+1}^{(b_1-1)\wedge s} \binom{n_1}{x_1} \binom{n_2}{s-x_1}$$

when $m = 2$. This p-value does not require specification of a either. We can reject H_0 and accept the study therapy when $\text{p-value} < \alpha^*$ regardless of the observed sample size at the stopping stage.

It is easy to show that the ordering of MLE depends on the critical values at the stopping stage. Hence, we can not calculate an MLE-based p-value if the sample size at the stopping stage is different from that specified by the design. Furthermore, the decision rule based on the MLE-based p-value may not match with that based on the two-stage testing rule whether the observed sample is identical to the planned one or not.

Example 4 (revisited): In Example 4, with 27 CRs out of $n = 30$ patients after the second stage, the p-value is 0.0086 by the above method for $p_0 = 0.7$ and $(a_1, n_1) = (4, 6)$. Since $\text{p-value} < \alpha^*$, this is another justification that Kim et al. [22] is a positive study. For $\alpha^* = 0.05$, this study will be positive with 26 CRs (p-value = 0.0259), but not with 25 CRs (p-value = 0.0605).

As mentioned above, the critical values of a two-stage design can not be used for testing if the realized sample size at the stopping stage is different from that specified at the design. In this case, we can conduct a statistical testing by calculating the p-value with the sample size at the stopping stage conditioned on the observed value and checking if it is smaller than the prespecified α level or not. In the previous section, we show that the Jennison-Turnbull confidence interval can be used for this purpose too. This property makes the p-value, together with the Jennison-Turnbull confidence interval, very useful for the analysis of two-stage phase II trials. Other p-value methods for multistage design [26,27] do not have these desirable properties.

Green and Dahlberg [13] and Herndon [14] have also considered under- or over-accrual of patients in each stage and proposed some ad hoc approaches to the selection of rejection values depending on the realized sample sizes. Chen and Ng [28] considered a set of combinations for possible sample sizes for stages 1 and 2 (n_1, n_2) and provide the rejection values (a_1, a) to minimize the maximal sample size or the average expected sample assuming that each combination in the set has the same probability to be observed. The latter approach seems to have issues: (i) If the combination of realized sample sizes does not belong to the prespecified set, it does not provide valid rejection values; (ii) For each combination of (n_1, n_2), rejection values (a_1, a) are given. For each n_1 value, there are multiple n_2 values with their own a_1 values. So, it is not clear which rejection value a_1 should be used for a realized n_1 after stage 1. Furthermore, suppose that we observed a combination (n_1, n_2) through two stages. Then, we will use a that is assigned to this combination, but a_1 that have been used after stage 1 may be different from the stage 1 rejection value assigned to this combination of sample sizes.

For $N_1 = n_1$ and $N_2 = n$, Chang and O'Brien [29] use the likelihood-ratio

$$\frac{f(m, s|\hat{p})}{f(m, s|p_0)} = \left(\frac{\hat{p}}{p_0}\right)^s \left(\frac{1-\hat{p}}{1-p_0}\right)^{N_m-s}$$

to measure how far the MLE $\hat{p} = \hat{p}(m, s)$ is from p_0 . Chang, Gould and Snapinn [30] and Cook [31] use the likelihood ratio ordering to derive p-values from sequential tests. Note that the ordering of the sample space based on the likelihood-ratio is two-sided in nature, so that it is inappropriate for p-values for phase II trials which have one-sided hypotheses. It is obvious from (3) that the UMVUE gives p-value satisfying the properties: (a) the p-values in the acceptance region of H_0 are larger than those in the rejection region, and (b) the p-value for the critical value matches with the type I error probability of the sequential testing. These properties are not satisfied by p-values derived from other orderings.

5 Discussions

Most popular design for phase II cancer clinical trials is single-arm two-stage design. The publications reporting these trials do not use appropriate statistical methods for data analysis. These trials are specified by sample sizes and critical values for two stages. Oftentimes the number of patients treated at each stage is different from that specified by the design. In this case, the critical values become meaningless. Probably due to this reason, publications reporting phase II trials are not clear if they accept or reject their study therapies, or some of them accept or reject them without a solid statistical ground.

We avoid this issue when analyzing a phase III clinical trial by conducting statistical analysis conditioning on the observed sample size. This is easier since phase III trials usually have large sample sizes and the statistical methods using large sample approximation do not require prespecification of a sample size. In the analysis of phase II trials with small sample sizes, we can similarly overcome this issue by conditioning on the realized sample size. This is a main theme of this review article. In this paper, we have

reviewed p-value and confidence interval methods for two-stage phase II clinical trials. The theory behind these methods is identical to that of two-stage testing of phase II clinical trials, so that any of these lead to the same conclusions if the sample size at the stopping stage is the same as that specified by the design. This is possible because all of these methods are based on a common theory called the UMVUE of RR. If the observed sample size for the stopping stage is different from that specified by the design, then the testing based on a two-stage critical values is of no use while we still can calculate unbiased p-value and confidence interval by extending the UMVUE calculation to this case.

We also have reviewed a search algorithm of two-stage phase II trial designs beyond the Simon's minimax and optimal designs that have been used for long by cancer biostatisticians. The new designs, called admissible designs, closely satisfy both the minimax and optimality criteria of Simon [2]. In this review article, we have limited our discussions to two-stage designs that are most popularly used for phase II cancer clinical trials, but all the methods discussed in this article can be extended to phase II trials with any number of stages.

Acknowledgments

This research was supported by a grant from the National Cancer Institute (CA142538-01).

References

1. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: Revised RECIST guideline (version 1.1). *European Journal of Cancer*. 2009; 45:228–247. [PubMed: 19097774]
2. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*. 1989; 10:1–10. [PubMed: 2702835]
3. Ludwig H, Viterbo L, Greil R, et al. Randomized phase II study of bortezomib, thalidomide, and dexamethasone with or without cyclophosphamide as induction therapy in previously untreated multiple myeloma. *Journal of Clinical Oncology*. 2013; 31:247–255. [PubMed: 23091109]
4. Friedberg JW, Mahadevan D, Cebula E, et al. Phase II study of alisertib, a selective aurora A kinase inhibitor, in relapsed and refractory aggressive B- and T-cell non-Hodgkin lymphomas. *Journal of Clinical Oncology*. 2014; 32:44–50. [PubMed: 24043741]
5. Koon HB, Krown SE, Lee JY, et al. Phase II trial of imatinib in AIDS-associated Kaposi sarcoma: AIDS alignancy Consortium Protocol 042. *Journal of Clinical Oncology*. 2014; 32:402–408. [PubMed: 24378417]
6. Jung SH, Carey M, Kim KM. Graphical search for two-stage phase II clinical trials. *Controlled Clinical Trials*. 2001; 22:367–372. [PubMed: 11514038]
7. Jung SH, Lee TY, Kim KM, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine*. 2004; 23:561–569. [PubMed: 14755389]
8. Chang MN, Therneau TM, Wieand HS, et al. Designs for group sequential phase II clinical trials. *Biometrics*. 1987; 43:865–874. [PubMed: 3427171]
9. Mander AP, Wason JMS, Sweeting MJ, Thompson SG. Admissible two-stage designs for phase II cancer clinical trials that incorporate the expected sample size under the alternative hypothesis. *Pharmaceutical Statistics*. 2012; 11:91–96. [PubMed: 22232071]
10. Englert S, Kieser M. Optimal adaptive two-stage designs for phase II cancer clinical trials. *Biometrical Journal*. 2013; 55:955–968. [PubMed: 23868324]
11. Kieser M, Englert S. Performance of Adaptive Designs for Single-Armed Phase II Oncology Trials. *Journal of Biopharmaceutical Statistics* 2014. 2014 Jun 6. Epub ahead of print.

12. Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine*. 2004; 23:881–896. [PubMed: 15027078]
13. Green SJ, Dahlberg S. Planned vs attained design in phase II clinical trials. *Statistics in Medicine*. 1992; 11:853–862. [PubMed: 1604065]
14. Herndon J. A design alternative for two-stage, phase II, multicenter cancer clinical trials. *Controlled Clinical Trials*. 1998; 19:440–450. [PubMed: 9741865]
15. Bowden J, Wason J. Identifying combined design and analysis procedures in two-stage trials with a binary end point. *Statistics in Medicine*. Published online 11 July 2012.
16. Jennison C, Turnbull BW. Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics*. 1983; 25:49–58.
17. Armitage P. Numerical studies in the sequential estimation of a binomial parameter. *Biometrika*. 1958; 45:1–15.
18. Tsiatis AA, Rosner GL, Mehta CR. Exact confidence intervals following a group sequential test. *Biometrics*. 1984; 40:797–803. [PubMed: 6518248]
19. Clopper CJ, Pearson ES. The use of confidence or fiducial limits illustrated in the case of the binomial. *Biometrika*. 1934; 26:404–413.
20. Duffy DE, Santner TJ. Confidence intervals for a binomial parameter based on multistage tests. *Biometrics*. 1987; 43:81–93.
21. Tsai WY, Chi Y, Chen CM. Interval estimation of binomial proportion in clinical trials with a two-stage design. *Statistics in Medicine*. 2008; 27:15–35. [PubMed: 17566141]
22. Kim SJ, Kim K, Kim BS, et al. Phase II trial of concurrent radiation and weekly cisplatin followed by VIPD chemotherapy in newly diagnosed, stage IE to IIE, nasal, extranodal NK/T-cell lymphoma: Consortium for Improving Survival of Lymphoma study. *Journal of Clinical Oncology*. 2009; 27:6027–6032. [PubMed: 19884539]
23. Shimada K, Suzuki R. Concurrent Chemoradiotherapy for Limited-Stage Extranodal Natural Killer/T-Cell Lymphoma, Nasal Type. *Journal of Clinical Oncology*. 2010; 28(14):e229. [PubMed: 20351319]
24. Kim SJ, Jung SH, Kim WS. Reply to K Shimada et al. *Journal of Clinical Oncology*. 2010; 28(14):e230.
25. Jung SH, Owzar K, George SL, et al. P-value calculation for multistage phase II cancer clinical trials (with discussion). *Journal of Biopharmaceutical Statistics*. 2006; 16:765–783. [PubMed: 17146978]
26. Cook TD. P-value adjustment in sequential clinical trials. *Biometrics*. 2002; 58:1005–1011. [PubMed: 12495156]
27. Koyama T, Chen H. Proper inferences from Simon's two-stage designs. *Statistics in Medicine*. 2008; 27:3145–3154. [PubMed: 17960777]
28. Chen TT, Ng TH. Optimal flexible designs in phase II clinical trials. *Statistics in Medicine*. 1998; 17:2301–2312. [PubMed: 9819829]
29. Chang MN, O'Brien PC. Confidence intervals following group sequential trials. *Controlled Clinical Trials*. 1985; 7:18–26. [PubMed: 3956209]
30. Chang MN, Gould AL, Snapinn SM. P-values for group sequential testing. *Biometrika*. 1995; 82:650–654.
31. Cook TD. P-value adjustment in sequential clinical trials. *Biometrics*. 2002; 58:1005–1011. [PubMed: 12495156]

Two-Stage Phase II Designs

Figure 1. Two-stage designs for $(p_0, p_1, \alpha^*, \beta^*) = (.1, .3, .05, .15)$ with $N = 37$

UMVUE, MLE, and probability mass function for true p for each observation in a two-stage design with $(a_1/n_1, a/n) = (3/13, 12/43)$

Table 1

m	s	$f(m, s, p)$ for p						
		UMVUE	MLE	0.1	0.2	0.3	0.4	0.5
1	0	0.000	0.000	0.254	0.055	0.010	0.001	0.000
1	1	0.077	0.077	0.367	0.179	0.054	0.011	0.002
1	2	0.154	0.154	0.245	0.268	0.139	0.045	0.010
1	3	0.231	0.231	0.100	0.246	0.218	0.111	0.035
2	4	0.308	0.093	0.001	0.000	0.000	0.000	0.000
2	5	0.312	0.116	0.004	0.002	0.000	0.000	0.000
2	6	0.317	0.140	0.007	0.006	0.001	0.000	0.000
	7	0.322	0.163	0.008	0.015	0.002	0.000	0.000
2	8	0.328	0.186	0.006	0.027	0.006	0.000	0.000
2	9	0.335	0.209	0.004	0.038	0.015	0.001	0.000
2	10	0.343	0.233	0.002	0.043	0.030	0.003	0.000
2	11	0.351	0.256	0.001	0.041	0.049	0.008	0.000
2	12	0.360	0.279	0.000	0.033	0.068	0.018	0.001
2	13	0.371	0.302	0.000	0.023	0.081	0.033	0.003
2	14	0.382	0.326	0.000	0.014	0.084	0.054	0.006
2	15	0.395	0.349	0.000	0.007	0.076	0.076	0.013
2	16	0.409	0.372	0.000	0.003	0.062	0.096	0.025
2	17	0.424	0.395	0.000	0.001	0.044	0.107	0.042
2	18	0.440	0.419	0.000	0.001	0.029	0.108	0.063
2	19	0.458	0.442	0.000	0.000	0.017	0.098	0.085
2	20	0.477	0.465	0.000	0.000	0.009	0.080	0.105
2	21	0.496	0.488	0.000	0.000	0.004	0.059	0.116
2	22	0.517	0.512	0.000	0.000	0.002	0.040	0.118
2	23	0.538	0.535	0.000	0.000	0.001	0.025	0.108
2	24	0.560	0.558	0.000	0.000	0.000	0.014	0.091
2	25	0.582	0.581	0.000	0.000	0.000	0.007	0.069
2	26	0.605	0.605	0.000	0.000	0.000	0.003	0.048

		$f(m, s d)$ for d						
m	s	UMVUE	MLE	0.1	0.2	0.3	0.4	0.5
2	27	0.628	0.628	0.000	0.000	0.000	0.001	0.030
2	28	0.651	0.651	0.000	0.000	0.000	0.001	0.017
2	29	0.674	0.674	0.000	0.000	0.000	0.000	0.009
2	30	0.698	0.698	0.000	0.000	0.000	0.000	0.004
2	31	0.721	0.721	0.000	0.000	0.000	0.000	0.002
2	32	0.744	0.744	0.000	0.000	0.000	0.000	0.001
2	33	0.767	0.767	0.000	0.000	0.000	0.000	0.000
2	34	0.791	0.791	0.000	0.000	0.000	0.000	0.000
2	35	0.814	0.814	0.000	0.000	0.000	0.000	0.000
2	36	0.837	0.837	0.000	0.000	0.000	0.000	0.000
2	37	0.861	0.861	0.000	0.000	0.000	0.000	0.000
2	38	0.884	0.884	0.000	0.000	0.000	0.000	0.000
2	39	0.907	0.907	0.000	0.000	0.000	0.000	0.000
2	40	0.930	0.930	0.000	0.000	0.000	0.000	0.000
2	41	0.954	0.954	0.000	0.000	0.000	0.000	0.000
2	42	0.977	0.977	0.000	0.000	0.000	0.000	0.000
2	43	1.000	1.000	0.000	0.000	0.000	0.000	0.000

Table 2
UMVUE and MLE, and p-value using these estimators for a two-stage design of $(n_1, n_2, a_1, b_1, a) = (34, 5, 17, 35, 20)$ with $p_0 = 0.4$

<i>m</i>	<i>s</i>	$f(m, s p_0)$	Estimate		p-value	
			UMVUE	MLE	UMVUE	MLE
1	0	0.0000	0.0000	0.0000	1.0000	1.0000
1	1	0.0000	0.0294	0.0294	1.0000	1.0000
1	2	0.0000	0.0588	0.0588	1.0000	1.0000
1	3	0.0001	0.0882	0.0882	1.0000	1.0000
1	4	0.0003	0.1176	0.1176	0.9999	0.9999
1	5	0.0010	0.1471	0.1471	0.9997	0.9997
1	6	0.0034	0.1765	0.1765	0.9986	0.9986
1	7	0.0090	0.2059	0.2059	0.9952	0.9952
1	8	0.0203	0.2353	0.2353	0.9862	0.9862
1	9	0.0391	0.2647	0.2647	0.9659	0.9659
1	10	0.0652	0.2941	0.2941	0.9268	0.9268
1	11	0.0948	0.3235	0.3235	0.8617	0.8617
1	12	0.1211	0.3529	0.3529	0.7669	0.7669
1	13	0.1366	0.3824	0.3824	0.6458	0.6458
1	14	0.1366	0.4118	0.4118	0.5092	0.5092
1	15	0.1214	0.4412	0.4412	0.3726	0.3726
1	16	0.0961	0.4706	0.4706	0.2512	0.2512
1	17	0.0679	0.5000	0.5000	0.1550	0.1550
2	18	0.0033	0.5294	0.4615	0.0872	0.2653
2	19	0.0129	0.5337	0.4872	0.0838	0.1517
2	20	0.0219	0.5403	0.5128	0.0709	0.1021
2	21	0.0217	0.5508	0.5385	0.0490	0.0559
2	22	0.0145	0.5672	0.5641	0.0273	0.0280
2	23	0.0075	0.5897	0.5897	0.0128	0.0128
2	24	0.0033	0.6154	0.6154	0.0053	0.0053
2	25	0.0013	0.6410	0.6410	0.0020	0.0020
2	26	0.0005	0.6667	0.6667	0.0007	0.0007

<i>m</i>	<i>s</i>	$f(m, s p_0)$	Estimate		p-value		
			UMVUE	MLE	UMVUE	MLE	Naive
2	27	0.0002	0.6923	0.6923	0.0002	0.0002	0.0002
2	28	0.0000	0.7179	0.7179	0.0001	0.0001	0.0001
2	29	0.0000	0.7436	0.7436	0.0000	0.0000	0.0000
2	30	0.0000	0.7692	0.7692	0.0000	0.0000	0.0000
2	31	0.0000	0.7949	0.7949	0.0000	0.0000	0.0000
2	32	0.0000	0.8205	0.8205	0.0000	0.0000	0.0000
2	33	0.0000	0.8462	0.8462	0.0000	0.0000	0.0000
2	34	0.0000	0.8718	0.8718	0.0000	0.0000	0.0000
2	35	0.0000	0.8974	0.8974	0.0000	0.0000	0.0000
2	36	0.0000	0.9231	0.9231	0.0000	0.0000	0.0000
2	37	0.0000	0.9487	0.9487	0.0000	0.0000	0.0000
2	38	0.0000	0.9744	0.9744	0.0000	0.0000	0.0000
2	39	0.0000	1.0000	1.0000	0.0000	0.0000	0.0000