Duke University
Duke Biostatistics and Bioinformatics (B&B) Working Paper Series

Year 2009 *Paper* 5

Two-Stage Phase II Clinical Trials with Heterogeneous Patient Populations

Sin-Ho Jung*

*sinho.jung@duke.edu

http://biostats.bepress.com/dukebiostat/art5

Two-Stage Phase II Clinical Trials with Heterogeneous Patient Populations

Sin-Ho Jung

Abstract

The patient population for a phase II trial often consists of multiple subgroups with different prognosis. In this case, a popular design approach is to specify the response rate and the prevalence of each subgroup, to calculate the response rate of the whole population by the weighted average of the response rates across subgroups, and to choose a standard phase II design such as Simon's optimal or minimax design to test on the response rate for the whole population. Although the prevalence of each subgroup is accurately specified, the observed prevalence among the accrued patients to the study may be quite different from the estimated one because of the small sample size, which is typical in most phase II trials. In this case, the fixed rejection value for a chosen standard phase II design may be either too conservative (i.e., increasing the false rejection probability of the experimental therapy) if the trial accrues more high-risk patients than expected or too anti-conservative (i.e., increasing the false acceptance probability of the experimental therapy) if the trial accrues more low-risk patients than expected. We can avoid such problem by adjusting the rejection value depending on the observed prevalence from the trial. In this paper, we investigate two flexible design approaches that choose rejection values depending on the observed prevalence, and compare them under various

Two-Stage Phase II Clinical Trials with Heterogeneous Patient Populations

Sin-Ho Jung [1]

SUMMARY

The patient population for a phase II trial often consists of multiple subgroups with different prognosis. In this case, a popular design approach is to specify the response rate and the prevalence of each subgroup, to calculate the response rate of the whole population by the weighted average of the response rates across subgroups, and to choose a standard phase II design such as Simon's optimal or minimax design to test on the response rate for the whole population. Although the prevalence of each subgroup is accurately specified, the observed prevalence among the accrued patients to the study may be quite different from the estimated one because of the small sample size, which is typical in most phase II trials. In this case, the fixed rejection value for a chosen standard phase II design may be either too conservative (i.e., increasing the false rejection probability of the experimental therapy) if the trial accrues more high-risk patients than expected or too anti-conservative (i.e., increasing the false acceptance probability of the experimental therapy) if the trial accrues more low-risk patients than expected. We can avoid such problem by adjusting the rejection value depending on the observed prevalence from the trial. In this paper, we investigate two flexible design approaches that choose rejection values depending on the observed prevalence, and compare them under various settings.

KEY WORDS: *Conditional power, Conditional type I error, Minimax design, Optimal design, Prevalence*

[1]Cancer and Leukemia Group B Statistical Center, and Department of Biostatistics and Bioinformatics, Duke University, Durham, North Carolina, 27710, U.S.A. (E-mail: sinho.jung@duke.edu)

1 Introduction

A single-arm phase II trial is to evaluate an experimental therapy compared to a historical control before it proceeds to a large scale phase III trial to be compared to a prospective control. The patient population for a phase II trial often consists of multiple subgroups with different prognosis. In this case, the final decision on the study treatment should adjust for the heterogeneity of the patient population.

Suppose that we want to evaluate the tumor response of CD30 antibody, SGN-30, combined with GVD (Gemcitabine, Vinorelbine, Pegylated Liposomal Doxorubicin) chemotherapy in patients with relapsed or refractory classical Hodgkin lymphoma (HL) through a phase II trial. In a previous study, GVD only has led to responses in 65% of patients with relapsed or refractory HL patients who never had a transplant and 75% in the transplant group. About 50% of patients in the previous study never had a transplant. Combining the data from the two cohorts, the response rate (RR) for the whole patient population is estimated as $p_0 = 70\%(= 0.5 \times 0.65 + 0.5 \times 0.75)$. When design a new phase II trial to evaluate the efficacy of the experimental therapy, one of the standard design accounting for the heterogeneity of the patient population is a single-arm trial based on the historical data to test

$$\tilde{H}_0 : p \leq 70\% \qquad \text{against} \qquad \tilde{H}_a : p > 70\%,$$

where p denotes the true RR of the experimental therapy in the patient population combining the two subgroups. By specifying a clinically significant RR, $p_1(> p_0)$, we may choose a standard phase II design, e.g. Simon's optimal or minimax design [1], or admissible design [2,3].

This approach can be biased if the observed proportion of patients from each subgroup is very different from the specified prevalence. London and Chang [4] resolve this issue by choosing rejection values depending on the observed prevalence. They propose to use early stopping boundaries for both low and high efficacy cases. A problem associated with this approach is that we reject or accept the experimental therapy for the whole population even though it may not be efficacious for a subgroup, so that the associated statistical testing may not be powerful in such cases. Pointing this out, Wathen et al. [5] propose a Bayesian method to test on the efficacy for each subgroup.

In this paper, we investigate phase II clinical trial design approaches based on above testing methods with early stopping for low efficacy only. For a testing on the whole population, we modify the London and Chang's method to allow for futility stopping only. For a design based on subgroup-specific testing, we propose a frequentist version of Wathen et al. [5]. We discuss two-stage designs for two subgroup cases only, but the proposed methods can be easily modified for multiple stage or multiple group cases. We compare their performance under various settings.

2 Two-Stage Designs

Because of ethical and economical issues, two-stage designs have been more popular for phase II cancer clinical trials than single-stage designs. We may stop a trial early when the RR of a study treatment turns out to be either too low or too high, but we consider the more general case with an early stopping due to a low RR only here. Under a two-stage design, we accrue n_k patients during stage $k(= 1, 2)$. Let $n = n_1 + n_2$. For stage $k(= 1, 2)$ and subgroup $j(= 1, 2)$, let m_{kj} and X_{kj} denote the number of patients and the number of responders, respectively. Note that $n_k = m_{k1} + m_{k2}$.

2.1 Design 1: Conventional Design based on a Specified Prevalence

Let p_j denote the RR of the experimental therapy for subgroup j. For subgroup j, we have null and alternative hypotheses, $H_{0j} : p_j = p_{0j}$ and $H_{aj} : p_j = p_{aj}$, respectively $(p_{0j} \leq p_{aj})$. We want to test $H_0 = H_{01} \cap H_{02}$ in favor of $H_a = H_{a1} \cup H_{a2}$. Given a specified prevalence of group 1, $r_1 (r_2 = 1 - r_1)$, let $p_0 = r_1 p_{01} + r_2 p_{02}$ and $p_a = r_1 p_{a1} + r_2 p_{a2}$. In order to test H_0 vs. H_a, we usually choose a two-stage design for $\tilde{H}_0 : p = p_0$ vs. $\tilde{H}_a; p = p_a$ such as Simon's minimax or optimal design, where p is the true RR for the whole population. Since the rejection values are fixed regardless of the observed prevalence, this approach can be biased if the observed number of patients from each subgroup is very different from the expected one based on the specified prevalence.

Example 1: For $(p_{01}, p_{02}) = (0.65, 0.75)$, $(p_{a1}, p_{a2}) = (0.8, 0.9)$, $r_1 = 0.5$, a type I error rate

of $\alpha^* = 0.1$ and a power of $1 - \beta^* = 0.9$, the Simon's minimax two-stage design to test

$$H_0 : p_0 = 0.7 \text{ against } H_a : p_1 = 0.85$$

proceeds as follows.

I. Stage 1: Accrue $n_1 = 22$ patients. If $\tilde{a}_1 = 15$ or fewer patients respond, then we stop the trial concluding that the combination therapy is inefficacious. Otherwise, the trial proceeds to stage 2.

II. Stage 2: Accrue an additional $n_2 = 30$ patients. If more than $\tilde{a} = 40$ patients out of the total $n = 52(= n_1 + n_2)$ respond, then the combination therapy will be accepted for further investigation.

We denote this design as $(n_1, n, a_1, a) = (22, 52, 15, 40)$.

If the trial rejects H_0, then we accept the experimental therapy for further investigation with respect to the whole population. Following two approaches use variable rejection values depending on the number of patients accrued from each subgroup.

2.2 Design 2: Based on Conditional Testing for the Whole Population

In this section, we consider a similar design setting as in Design 1, but we do not need an accurate specification of the prevalence of each subgroup by choosing rejection values depending on the observed prevalence.

At first, we choose sample sizes for two stages, n_1 and n_2. We may choose them from a conventional two-stage design based on a projected prevalence as discussed in the previous section. Note that the projected prevalence does not have to be the true one, and an erroneously chosen one does not have any impact on the validity (i.e., type I error rate) of the chosen two-stage design that will be investigated in this section.

Among n_k patients entered during stage k, let m_{kj} denote the number of patients from subgroup j. Given n_k and m_{kj} for $k = 1, 2$ and $j = 1, 2$, we propose to choose the stage 1

rejection value $a_1 = [m_{11}p_{01} + m_{12}p_{02}]$, where $[c]$ denotes the largest integer not exceeding c. In other words, we reject the experimental therapy early if the observed number of responders from stage 1 is no larger than the expected number of responders under H_0. The stage 1 rejection values of the Simon's designs with homogeneous patient populations satisfy this condition.

Given a type I error rate α^*, the rejection value of the second stage a is the largest integer satisfying $\alpha \leq \alpha^*$, where

$$\begin{aligned}\alpha &= \mathrm{P}(X_{11} + X_{12} > a_1, X_{11} + X_{12} + X_{21} + X_{22} > a | p_{01}, p_{02}, m_{11}, m_{21}) \\ &= \sum_{x_{11}=0}^{m_{11}} \sum_{x_{12}=0}^{m_{12}} \sum_{x_{21}=0}^{m_{21}} \sum_{x_{22}=0}^{m_{22}} I(x_{11} + x_{12} > a_1, x_{11} + x_{12} + x_{21} + x_{22} > a) \\ &\quad \times b(x_{11}|m_{11}, p_{01}) b(x_{12}|m_{12}, p_{02}) b(x_{21}|m_{21}, p_{01}) b(x_{22}|m_{22}, p_{02}) \qquad (1)\end{aligned}$$

where

$$b(x|n,p) = \binom{n}{x} p^x (1-p)^{n-x} \text{ for } x = 0, 1, ..., n$$

denotes the probability mass function of the binomial distribution with number of independent trials n and success probability p.

Given $(n_1, n_2, m_{11}, n_{21}, a_1, a)$, the conditional power under H_a is given as

$$\begin{aligned}1 - \beta &= \mathrm{P}(X_{11} + X_{12} > a_1, X_{11} + X_{12} + X_{21} + X_{22} > a | p_{a1}, p_{a2}) \\ &= \sum_{x_{11}=0}^{m_{11}} \sum_{x_{12}=0}^{m_{12}} \sum_{x_{21}=0}^{m_{21}} \sum_{x_{22}=0}^{m_{22}} I(x_{11} + x_{12} > a_1, x_{11} + x_{12} + x_{21} + x_{22} > a) \\ &\quad \times b(x_{11}|m_{11}, p_{a1}) b(x_{12}|m_{12}, p_{a2}) b(x_{21}|m_{21}, p_{a1}) b(x_{22}|m_{22}, p_{a2}).\end{aligned}$$

In summary, a two-stage phase II trial based on the conditional testing for the whole population is carried out as follows.

I. Specify design parameters: $\{(p_{0j}, p_{aj}), j = 1, 2\}$, $(\alpha^*, 1 - \beta^*)$, prevalence r_1.

II. Choose n_1 and n_2 based on the design parameters.

Stage 1: Accrue n_1 patients.

(a) Given the observed number of patients m_{1j} for $j = 1, 2$, calculate $a_1 = [m_{11}p_{01} + m_{12}p_{02}]$.

(b) If the number of responders $X_1 = X_{11} + X_{12}$ is smaller than or equal to a_1, stop the trial. Otherwise, proceed to stage 2.

Stage 2: Accrue an additional n_2 patients.

(a) Given the observed number of patients m_{kj} for $k = 1, 2$ and $j = 1, 2$, find the maximum value of a satisfying $\alpha = \alpha(m_{11}, m_{21}, a) \leq \alpha^*$ by (1).

(b) If the cumulative number of responders $X = X_1 + X_2$ is larger than a, then accept the study therapy for further investigation.

As in Design 1, rejection of H_0 by Design 2 means accepting the study therapy for the whole population.

Example 2: For $(p_{01}, p_{02}) = (0.65, 0.75)$ and $(\alpha^*, 1-\beta^*) = (0.1, 0.9)$ considered in Example 1, suppose that we choose $n_1 = 22$ and $n_2 = 30$ based on a projected prevalence $r_1 = 0.5$. Table 1 lists the rejection values (a_1, a) and the associated $(\alpha, 1 - \beta)$ for some chosen outcomes of (m_{11}, m_{21}). Note that if the observed prevalence rate from each stage, m_{11}/n_1 and m_{21}/n_2, is close to the projected $r_1 = 0.5$, then the rejection values are the same as those of the Simon's minimax design, $(a_1, a) = (15, 40)$, from Example 1. However, if the observed prevalence, m_{11}/n_1 or m_{21}/n_2, is lower (higher) than the projected $r_1 = 0.5$, then the rejection values become larger (smaller) than those of the minimax design. While the conditional type I error rate α is controlled below $\alpha^* = 0.1$, the conditional power is maintained around the desired $1 - \alpha^* = 0.9$ level.

(Table 1 may be placed around here.)

London and Chang [4] propose two stage-designs with both upper and lower stopping values obtained conditioning on the observed prevalence. The lower (corresponding to our rejection value a_1) and upper stopping values at stage 1 are chosen by assigning a fraction of the total type I error rate α^* and power $1 - \beta^*$, respectively. They choose n_1 and n_2 to minimize the maximal sample size $n = n_1 + n_2$ among the two-stage designs satisfying the type I error rate and power condition when m_{1j} and m_{2j} are close to the expected numbers for

a specified prevalence. On the other hand, there being no compelling ethical argument and thus rarely used, we consider early stopping only for lack of efficacy in this paper. Further, we choose n_1 and n_2 following the concept of Simon's [1] designs for phase II trials with lower stopping boundary only.

2.3 Design 3: Based on Conditional Testing for each Subgroup

Above phase II trial designs accept or reject the study therapy for the whole population. There may exist cases where a new therapy is efficacious for one subgroup, but not for the other. In this case, above designs may have a low power. Pointing this out, Wathan et al. [5] propose to test the experimental therapy with respect to each subgroup using Bayesian method. In this section, we propose a two-stage design method based on frequentist hypothesis testing.

As in Design 2, we choose sample sizes in two stages, n_1 and n_2, from a conventional design of Section 2.1 based on a projected prevalence. At the end of stage 1, suppose that m_{1j} patients are accrued from subgroup j ($m_{11} + m_{12} = n_1$). We reject the experimental therapy for subgroup j if the number of responders is smaller than or equal to the expected number under H_{0j}, i.e. $X_{1j} \leq a_{1j}$ with $a_{1j} = [m_{1j}p_{0j}]$. We also close a subgroup to patient accrual if it does not accrue any patient during stage 1. Even if we leave this subgroup open in the second stage, we may not have enough number of patients to properly evaluate the study therapy for this subgroup. The study proceeds to the second stage to enter the patients only to the subgroups that survive the first stage. The second stage accrue n_2 patients as planned whether a subgroup is dropped after the first stage or not.

The stage 2 rejection values a_j for subgroup j is determined to control a chosen false positivity. The false positivity corresponding to the type I error rate α of Designs 1 and 2 is the trial-wise error rate (TWER) defined as the probability of accepting the study therapy for any subgroup under H_0. Suppose that we use an equal subgroup-specific type I error γ for each subgroup, i.e. $\gamma = \mathrm{P}(\text{accept the therapy for subgroup } j|H_{0j})$. Since the decisions on two subgroups are independent, we have $\alpha = 1 - (1-\gamma)^2$. Hence, in order to control the TWER at α^*, we have to control the subgroup-specific type I error rate at $\gamma^* = 1 - \sqrt{1-\alpha^*}$. When subgroup j proceeds to the second stage, we accept the study therapy for this subgroup

if the cumulative number of responders $X_{1j} + X_{2j}$ is larger than a_j, where a_j is the largest integer satisfying $\gamma_j \leq \gamma^*$, where $\gamma_j = \gamma_j(m_{1j}, m_{2j})$ is the subgroup-specific type I error rate defined as

$$\begin{aligned}\gamma_j &= \mathrm{P}(X_{1j} > a_{1j}, X_{1j} + X_{2j} > a_j | p_{0j}, m_{1j}, m_{2j}) \\ &= \sum_{x_{1j}=a_{1j}+1}^{m_{1j}} \sum_{x_{2j}=0}^{m_{2j}} I(x_{1j} + x_{2j} > a_j) b(x_{1j}|m_{1j}, p_{0j}) b(x_{2j}|m_{2j}, p_{0j}). \end{aligned} \quad (2)$$

Note that if $X_{1,3-j} \leq a_{1,3-j}$, then we have $m_{2j} = n_2$ and $m_{2,3-j} = 0$. Consequently, a_j as well as m_{2j} depend on the number of responders in the other subgroup $X_{1,3-j}$.

In fact, (2) is the subgroup-specific type I error conditioning on the event $\{X_{1,3-j} \leq a_{1,3-j}\}$. The conditioning event on subgroup $3-j$ is theoretically independent of the event on the main event on subgroup j, but the conditional probability changes slightly depending on the conditioning because of the discreteness of binomial distributions. Let a'_j and a''_j denote the stage 2 rejection value for subgroup j under $\{X_{1,3-j} > a_{1,3-j}\}$ and $\{X_{1,3-j} \leq a_{1,3-j}\}$, respectively. Let A_j denote the event that we accept the study therapy for subgroup j. Then, given the observed number of patients from subgroup 1 (m_{11}, m_{21}) when both subgroups survive over stage 1, the exact TWER is calculated as

$$\begin{aligned}\alpha &= \mathrm{P}(A_1|H_0, m_{11}, m_{21}) + \mathrm{P}(A_2|H_0, m_{11}, m_{21}) - \mathrm{P}(A_1 \cap A_2|H_0, m_{11}, m_{21}) \\ &= \sum_{j=1}^{2} \{\mathrm{P}(X_{1,3-j} > a_{1,3-j}|p_{0,3-j}, m_{1,3-j}) \mathrm{P}(X_{1j} > a_{1j}, X_{1j} + X_{2j} > a'_j | p_{0j}, m_{1j}, m_{2j}) \\ &\quad + \mathrm{P}(X_{1,3-j} \leq a_{1,3-j}|p_{0,3-j}, m_{1,3-j}) \mathrm{P}(X_{1j} > a_{1j}, X_{1j} + X_{2j} > a''_j | p_{0j}, m_{1j}, m_{2j} = n_2)\} \\ &\quad - \prod_{j=1}^{2} \mathrm{P}(X_{1j} > a_{1j}, X_{1j} + X_{2j} > a'_j | p_{0j}, m_{1j}, m_{2j}) \\ &= \sum_{j=1}^{2} \{\bar{B}(a_{1,3-j}|m_{1,3-j}, p_{0,3-j}) \sum_{x_{1j}=a_{1j}+1}^{m_{1j}} \sum_{x_{2j}=0}^{m_{2j}} I(x_{1j} + x_{2j} > a'_j) b(x_{1j}|m_{1j}, p_{0j}) b(x_{2j}|m_{2j}, p_{0j}) \\ &\quad + B(a_{1,3-j}|m_{1,3-j}, p_{0,3-j}) \sum_{x_{1j}=a_{1j}+1}^{m_{1j}} \sum_{x_{2j}=0}^{n_2} I(x_{1j} + x_{2j} > a''_j) b(x_{1j}|m_{1j}, p_{0j}) b(x_{2j}|n_2, p_{0j})\} \\ &\quad - \prod_{j=1}^{2} \{\sum_{x_{1j}=a_{1j}+1}^{m_{1j}} \sum_{x_{2j}=0}^{m_{2j}} I(x_{1j} + x_{2j} > a'_j) b(x_{1j}|m_{1j}, p_{0j}) b(x_{2j}|m_{2j}, p_{0j})\}, \end{aligned}$$

where $B(x|n,p) = \sum_{i=0}^{x} b(i|n,p)$ and $\bar{B}(x|n,p) = 1 - B(x|n,p)$. The trial-wise power is calculated similarly under H_a.

In summary, a two-stage phase II trial based on a conditional testing for each subpopulation is conducted as follows.

I. Specify design parameters: $\{(p_{0j}, p_{aj}), j = 1, 2\}$, $(\alpha^*, 1 - \beta^*)$ and r_1.

II. Choose n_1 and n_2 based on the design parameters.

III. Stage 1: Accrue n_1 patients.

(a) For subgroup j, calculate $a_{1j} = [m_{1j}p_{0j}]$.

(b) If the number of responders X_{1j} is larger than a_{1j}, then subgroup j proceeds to stage 2. Otherwise, subgroup j is closed. If both subgroups are closed, the whole trial is stopped.

IV. Stage 2: Enter an additional n_2 patients to the subgroups that survive the first stage.

(a) For subgroup j, find the largest integer a_j satisfying $\gamma_j = \gamma_j(m_{1j}, m_{2j}, a_j) \leq 1 - \sqrt{1 - \alpha^*}$ from (2).

(b) If the cumulative number of responders $X_{1j} + X_{2j}$ is larger than a_j, then accept the study therapy for further investigation in subgroup j.

If the experimental therapy is rejected for the larger subgroup after the first stage, the investigators may consider closing the whole study since the study will take a long time to accrue n_2 patients from a small subpopulation during the second stage and a therapy efficacious only for a small subgroup is less interesting.

2.4 Unconditional Type I Error Rate and Power

If the true prevalence of subgroup 1 is r_1, then m_{11} and m_{21} are independent binomial random variables with 'success' probability r_1 and number of trials n_1 and n_2, respectively [4]. Given (m_{11}, m_{21}), let $\alpha = \alpha(m_{11}, m_{21})$ and $1 - \beta = 1 - \beta(m_{11}, m_{21})$ denote the conditional type I error rate and power, respectively, for one of the design methods discussed above. We can calculate the unconditional type I error and power of the design method by

$$\bar{\alpha} = \sum_{m_{11}=0}^{n_1} \sum_{m_{21}=0}^{n_2} \alpha(m_{11}, m_{21}) b(m_{11}|n_1, r_1) b(m_{21}|n_2, r_1)$$

$$1 - \bar{\beta} = \sum_{m_{11}=0}^{n_1} \sum_{m_{21}=0}^{n_2} \{1 - \beta(m_{11}, m_{21})\} b(m_{11}|n_1, r_1) b(m_{21}|n_2, r_1).$$

For example, if the true prevalence is $r_1 = 0.5$ for the lymphoma trial discussed in the introduction, the Simon's minimax design (Design 1) has $(\bar{\alpha}, 1-\bar{\beta}) = (0.0980, 0.9029)$ which are identical to the type I error rate and power with respect to hypotheses

$$\tilde{H}_0 : p_0 = 0.7 \text{ against } \tilde{H}_a : p_1 = 0.85.$$

Based on $r_1 = 0.5$, Designs 2 and 3 have $(\bar{\alpha}, 1-\bar{\beta}) = (0.0772, 0.8825)$ and $(0.06812, 0.7775)$. Note that $\bar{\alpha}$ of Design 2 is smaller than that of Design 1 since Design 2 controls the type I error below $\alpha^* = 0.1$ for all values of $m_{11}(\in [0, n_1])$ and $m_{21}(\in [0, n_2])$ while Design does so only for $m_{11} \approx r_1 n_1$ and $m_{21} \approx r_1 n_2$. Design 3 is slightly more conservative than Design 2 because it controls the conditional type I errors within each subgroup.

3 Numerical Studies

In this chapter, we compare the performance of the three design methods for two real phase II clinical trials with two subgroups.

At first, we investigate the phase II trial of CD30 antibody, SGN-30, combined with GVD in patients with relapsed or refractory classical HL that was briefly introduced in Section 1. For $(\alpha^*, 1-\beta^*) = (0.1, 0.9)$ and $(p_{01}, p_{02}, p_{a1}, p_{a2}) = (0.65, 0.75, 0.8, 0.9)$, we choose the Simon's minimax design $(n_1, n_2, a_1, a) = (22, 30, 15, 40)$ as Design 1, and $(n_1, n_2) = (22, 30)$ for the stages 1 and 2 sample sizes of Designs 2 and 3. Figure 1(a) displays the conditional type I error rates $\alpha = \alpha(m_{11}, m_{21})$ (the lower three lines) and powers $1-\beta = 1-\beta(m_{11}, m_{21})$ (the upper three lines) of the three designs. The blue lines are for Design 1, the red lines are for Design 2 and the green lines are for Design 3. Only m_{11} values are marked in the x-axis, but actually m_{21} values run from 0 to $n_2 = 30$ between consecutive m_{11} values. Consequently, the conditional type I error and power, especially for Design 1, regularly fluctuate between consecutive m_{11} values.

We observe that Design 1 has conditional type I error rate fluctuating between 0.021 and 0.306, while Designs 2 and 3 tightly control it below the specified $\alpha^* = 0.1$. The conditional power of Design 2 is closely maintained around $1-\beta^* = 0.1$, but that of Design 1 widely changes between 0.635 and 0.992 and that of Design 3 is underpowered (between 0.648 and 0.893) over the range of m_{11} and m_{22} values.

Table 1(b) displays the conditional powers when the study therapy is efficacious only for subgroup 1, i.e. $H_a : p_{a1} = 0.8, p_{a2} = 0.75$. We observe that Design 3 has a higher conditional power than Design 2 for a wide range of (m_{11}, m_{21}) values. We also observe that the conditional powers for Designs 2 and 3 tend to increase in m_{11} and m_{21} since large m_{11} and m_{21} values mean that most patients are recruited from subgroup 1 for which the study therapy is efficacious. If m_{11} and m_{21} are small these designs have very low powers.

Table 1(c) displays the conditional powers when the study therapy is efficacious only for subgroup 2, i.e. $H_a : p_{a1} = 0.65, p_{a2} = 0.9$. We observe that Design 3 has a higher conditional power than Design 2 for a wide range of (m_{11}, m_{21}) values in this case too. The conditional powers for Designs 2 and 3 tend to decrease in m_{11} and m_{21} since large m_{11} and m_{21} values mean that most patients are recruited from subgroup 2 for which the study therapy is efficacious.

Wathen et al. [5] consider a phase II trial for AML patients who relapsed after an initial CR complete remission (CR). From historical data, a standard therapy as a 'salvage' therapy is shown to have a response rate (RR) of $p_{01} = 0.1$ for the patients whose first CR duration was shorter than 1 year (high-risk group, called subgroup 1), and a RR of $p_{02} = 0.45$ for the patients whose first CR duration is longer than 1 year (low-risk group, called subgroup 2). The goal of the study is to determine if the low-risk group increases the RR to $p_{a1} = 0.3$ and the high-risk group increases the RR to $p_{a2} = 0.6$. The historical data include $r_1 = 70\%$ of high-risk patients. Combining the data from the two subgroups, the RR for the whole population can be specified as $p_0 = 0.205 (= 0.7 \times 0.1 + 0.3 \times 0.45)$ under H_0 and $p_a = 0.39 (= 0.7 \times 0.3 + 0.3 \times 0.6)$ under H_a. For $(\alpha^*, 1 - \beta^*) = (0.1, 0.9)$, the Simon's optimal design (Design 1) is given as $(n_1, n, a_1, a) = (20, 45, 4, 12)$. We also choose $(n_1, n_2) = (20, 25)$ for Designs 2 and 3.

Figure 2(a) displays the conditional type I error rates and powers of the three designs. Since $r_1 = 0.7$ is much larger than the middle point 0.5, the conditional type I error rate of Design 1 wildly fluctuates (between 0.000 and 0.903) in a wide range of small m_{11} and m_{21} values. If m_{11}/n_1 and m_{21}/n_2 are much smaller than $r_1 = 0.7$, then this trial accrued much more patients from the low-risk group (subgroup 2) than expected. Consequently, the fixed rejection values of Design 1 will be too anti-conservative in this case. In contrast, Designs 2 and 3 with variable rejection values depending on the observed prevalence tightly control

the conditional type I error below $\alpha^* = 0.1$. Designs 2 and 3 have similar conditional powers overall, but the former is slightly more powerful for m_{11} values smaller than 5.

Table 2(b) displays the conditional powers when the study therapy is efficacious only for subgroup 1, i.e. $H_a : p_{a1} = 0.3, p_{a2} = 0.45$. We observe that Design 3 has a higher conditional power than Design 2 for a wide range of (m_{11}, m_{21}) values. We also observe that the conditional powers for Designs 2 and 3 tend to increase in m_{11} and m_{21} since large m_{11} and m_{21} values mean that most patients are recruited from subgroup 1 for which the study therapy is efficacious. However, the conditional power of Design 1 has a reverse trend because of the sharply decreasing trend of the conditional type 1 error rate.

Table 2(c) displays the conditional powers when the study therapy is efficacious only for subgroup 2, i.e. $H_a : p_{a1} = 0.1, p_{a2} = 0.6$. Designs 2 and 3 have similar conditional powers, but the latter is slightly more powerful when the observed prevalence is close to the expected one, i.e. $m_{11} \approx r_1 n_1$ and $m_{21} \approx r_1 n_2$. The conditional powers for Designs 2 and 3 tend to decrease in m_{11} and m_{21} since large m_{11} and m_{21} values mean that most patients are recruited from subgroup 2 for which the study therapy is efficacious.

4 Discussions

As the associate editor points out, patient heterogeneity is an immensely important problem that arises frequently in most clinical trials. Accounting for the randomness of observed subgroup sample sizes arises in any test where patients are heterogeneous, so this issue is not limited to phase II. While this issue is well addressed in phase III clinical trials by stratified randomization, the issue in phase II trials that are mostly conducted as single-arm trials has been widely ignored. We can avoid such issue by randomizing patients between the experimental arm and a prospective control [6-8] as in phase III trials. Adopting a stratified randomization, a randomized phase II trial definitely guarantees an unbiased comparison, but it requires up to 4 times larger sample size compared to a single-arm trial.

When the historical control data come from a small prior study, we may use the estimates from the prior study as the design parameters p_{0j}, but we should incorporate the variation of these control parameters in designing a new study [9,10]. If the RR for the control is very

low, e.g. 0.05 or 0.1, then we may be comfortable with a single-arm trial even with design parameters with some variation. However, for stable diseases with a larger RR, we may consider a randomized phase II trial when the estimated design parameters are unreliable.

In this paper, we assume that there exist reliable historical data for each subgroup of a study population as in a standard single-arm phase II trial case. Under the assumption, we have studied two flexible designs for single-arm phase II trials with heterogeneous patient populations. Each design chooses rejection values to accurately control the conditional type I error depending on the number of patients accrued from different subgroups during each stage. The first design (Design 2) rejects or accepts the study therapy for the whole population, whereas the second one (Design 3) rejects or accepts the study therapy for each subgroup. Through real examples, Design 2 is shown to be slightly more powerful than Design 3 when the study therapy is efficacious for both subgroups. However, if therapy is efficacious only for one subgroup, then Design 3 is shown to be more powerful.

The proposed designs can be easily extended to the cases with more than two subgroups. A Fortran program for these designs are available from the author.

ACKNOWLEDGEMENTS

The author wishes to express his sincere thanks to Dr. Richard Simon for reading the manuscript and making helpful suggestions.

REFERENCES

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; 10: 1-10.

2. Jung SH, Carey M and Kim KM. Graphical search for two-stage phase II clinical trials. *Controlled Clinical Trials* 2001; 22: 367-372.

3. Jung SH, Lee TY, Kim KM, George S. Admissible two-stage designs for phase II cancer clinical trials. *Statistics in Medicine* 2004; 23: 561-569.

4. London WB, Chang MN. One- and two-stage designs for stratified phase II clinical trials. *Statistics in Medicine* 2005; 24: 2597-2611.

5. Wathen JK, Thall PF, Cook JD, Estey EH. Accounting for patient heterogeneity in phase II clinical trials. *Statistics in Medicine* 2008; 27: 2802-2815.

6. Thall PF, Simon R, Ellenberg SS. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* 1989; 45: 537-547.

7. Rubinstein LV, Korn EL, Freidlin B, Hunsberger S, Ivy SP, Smith MA. Design issues of randomized phase II trials and a proposal for phase II screening trials. *Journal of Clinical Oncology* 2005; 23(28): 7199-7206.

8. Jung SH. Randomized phase II trials with a prospective control. *Statistics in Medicine* 2008; 27: 568-583.

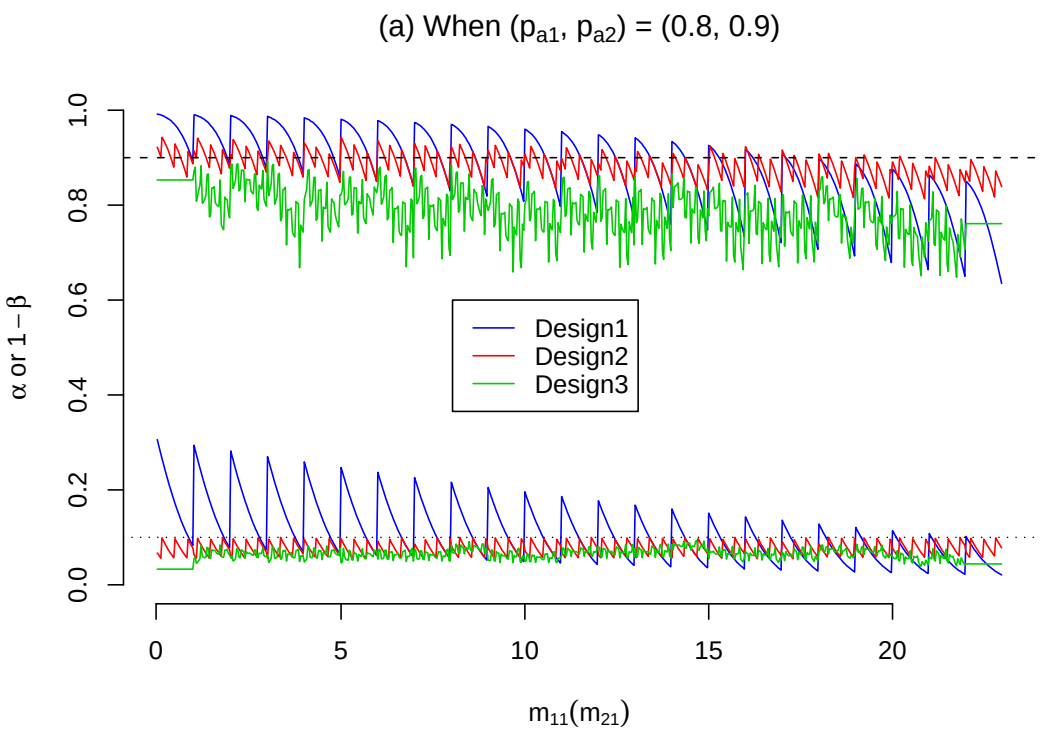
9. Makuch RW, Simon RM. Sample size considerations for non-randomized comparative studies. *Journal of Chronic Disease* 1980; 33: 175-181.

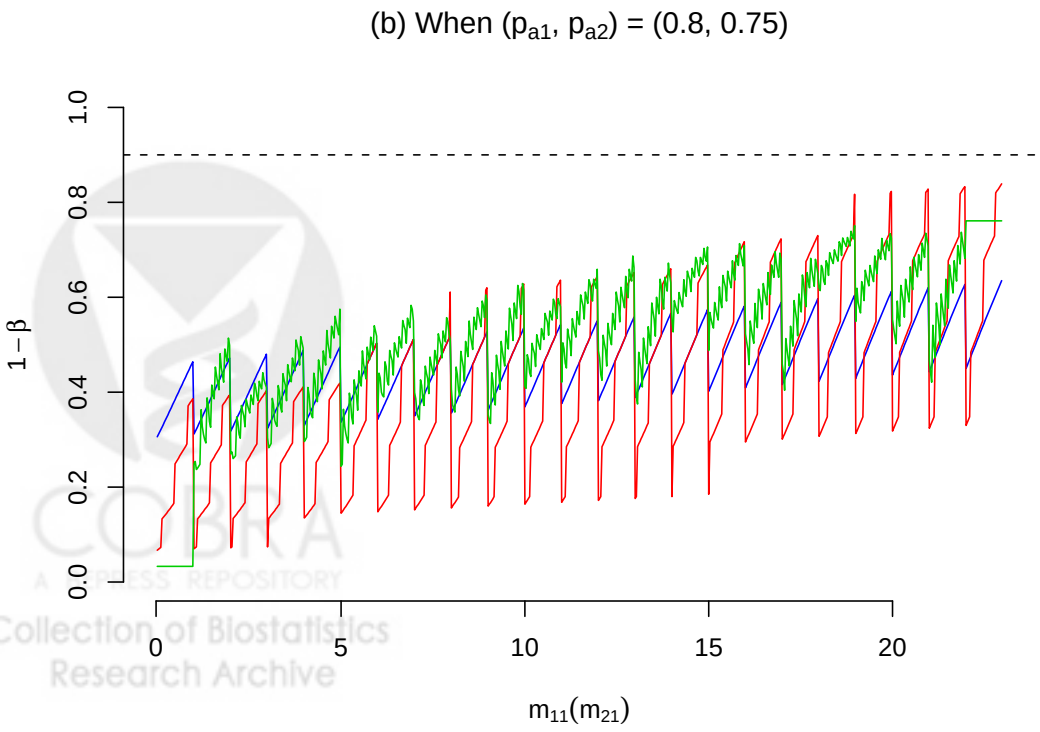
10. Thall PF, Simon R. Incorporating historical control data in planning phase II clinical trials. *Statistics in Medicine* 1990; **9**: 215-228.

Table 1. Design 2 with $(n_1, n_2) = (22, 30)$ under $(p_{01}, p_{02}, p_{a1}, p_{a2}, \alpha^*, 1-\beta^*) = (0.65, 0.75, 0.8, 0.9, 0.1, 0.9)$ and various (m_{11}, m_{21}) values

(m_{11}, m_{21})	(a_1, a)	$(\alpha, 1-\beta)$	(m_{11}, m_{21})	(a_1, a)	$(\alpha, 1-\beta)$
$(7, 6)$	$(15, 42)$	$(0.061, 0.890)$	$(11, 18)$	$(15, 40)$	$(0.083, 0.888)$
$(7, 9)$	$(15, 41)$	$(0.095, 0.925)$	$(11, 21)$	$(15, 40)$	$(0.071, 0.867)$
$(7, 12)$	$(15, 41)$	$(0.081, 0.906)$	$(11, 24)$	$(15, 40)$	$(0.061, 0.842)$
$(7, 15)$	$(15, 41)$	$(0.069, 0.885)$	$(13, 6)$	$(15, 41)$	$(0.079, 0.893)$
$(7, 18)$	$(15, 41)$	$(0.058, 0.860)$	$(13, 9)$	$(15, 41)$	$(0.067, 0.873)$
$(7, 21)$	$(15, 40)$	$(0.090, 0.902)$	$(13, 12)$	$(15, 40)$	$(0.100, 0.904)$
$(7, 24)$	$(15, 40)$	$(0.077, 0.881)$	$(13, 15)$	$(15, 40)$	$(0.086, 0.888)$
$(9, 6)$	$(15, 41)$	$(0.099, 0.926)$	$(13, 18)$	$(15, 40)$	$(0.074, 0.868)$
$(9, 9)$	$(15, 41)$	$(0.085, 0.909)$	$(13, 21)$	$(15, 40)$	$(0.063, 0.846)$
$(9, 12)$	$(15, 41)$	$(0.072, 0.890)$	$(13, 24)$	$(15, 39)$	$(0.094, 0.883)$
$(9, 15)$	$(15, 41)$	$(0.061, 0.866)$	$(15, 6)$	$(14, 41)$	$(0.074, 0.893)$
$(9, 18)$	$(15, 40)$	$(0.093, 0.905)$	$(15, 9)$	$(14, 41)$	$(0.062, 0.869)$
$(9, 21)$	$(15, 40)$	$(0.080, 0.885)$	$(15, 12)$	$(14, 40)$	$(0.096, 0.909)$
$(9, 24)$	$(15, 40)$	$(0.068, 0.863)$	$(15, 15)$	$(14, 40)$	$(0.082, 0.889)$
$(11, 6)$	$(15, 41)$	$(0.089, 0.910)$	$(15, 18)$	$(14, 40)$	$(0.070, 0.866)$
$(11, 9)$	$(15, 41)$	$(0.076, 0.892)$	$(15, 21)$	$(14, 40)$	$(0.059, 0.840)$
$(11, 12)$	$(15, 41)$	$(0.064, 0.871)$	$(15, 24)$	$(14, 39)$	$(0.090, 0.885)$
$(11, 15)$	$(15, 40)$	$(0.097, 0.906)$	$(15, 27)$	$(14, 39)$	$(0.077, 0.863)$

Figure 1: Conditional type I error and power of two-stage designs under $(\alpha^*, 1-\beta^*) = (0.1, 0.9)$ and $(p_{01}, p_{02}) = (0.65, 0.75)$. The blue, green and red lines are for the designs based on conventional, whole population and subgroup-specific testing, and the upper three lines are conditional powers and the lower three lines are conditional type I error.





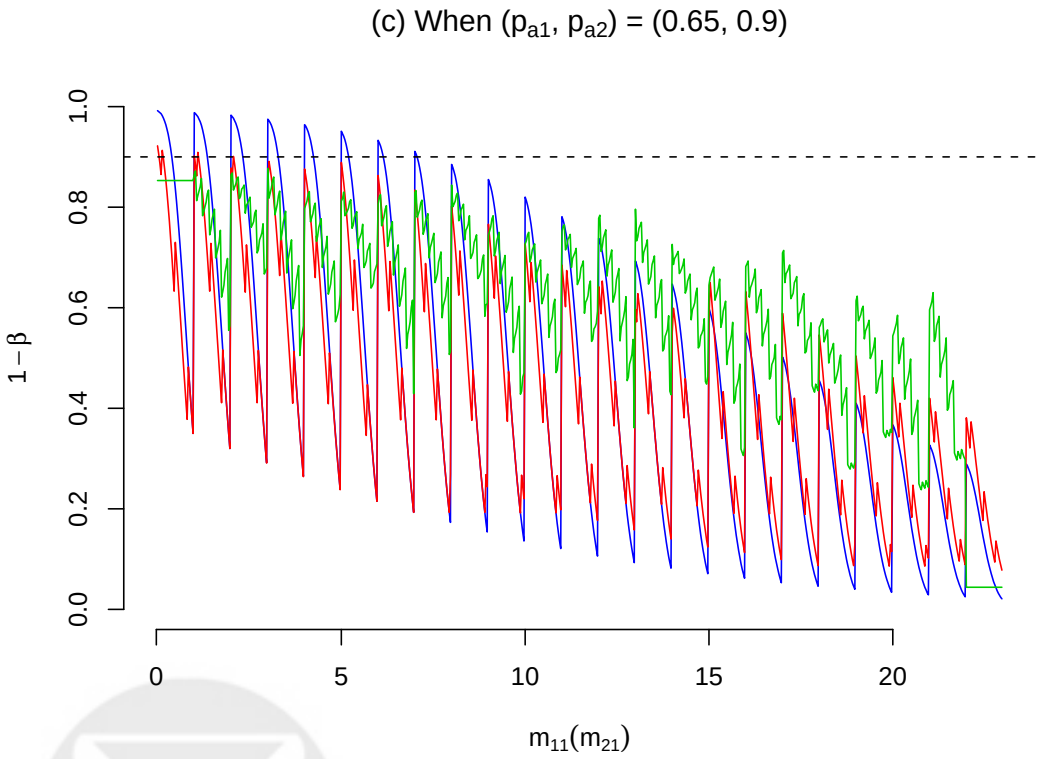


Figure 2: Conditional type I error and power of two-stage designs under $(\alpha^*, 1-\beta^*) = (0.1, 0.9)$ and $(p_{01}, p_{02}) = (0.1, 0.45)$. The blue, green and red lines are for the designs based on conventional, whole population and subgroup-specific testing, and the upper three lines are conditional powers and the lower three lines are conditional type I error.