Published in final edited form as:

J Biopharm Stat. 2009 ; 19(3): 456–468. doi:10.1080/10543400902802391.

Between-Arm Comparisons in Randomized Phase II Trials

Sin-Ho Jung¹ and Stephen L. George

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

SUMMARY

In a phase II trial, we may randomize patients to multiple arms of experimental therapies and evaluate their efficacy to determine if any of them is worth of a large scale phase III trial. Usually the primary objective of such study is to identify experimental therapies that are efficacious compared to a historical control. Each arm is independently evaluated using a standard design for single-arm phase II trial, e.g. Simon's optimal or minimax design. When more than one arms are accepted through such a randomized trial, we may want to select the winner(s) among them. There are methods for between-arm comparisons in the literature, but most of them have drawbacks. They have a large false selection probability (type I error) when the competing arms have a small difference in efficacy, or the statistical tests used in the selection procedure do not properly reflect the small sample sizes and multi-stage design of the trials. In this paper, we propose between-arm comparison methods for selection in randomized phase II trials addressing these issues.

Keywords

Pick-the-winner; Type I error; Type II error; Two-stage design; Uniformly minimum variance unbiased estimator

1 Introduction

Phase II clinical trials are designed to screen out experimental therapies with low efficacy before they proceed to a large scale phase III trial. Often, we have multiple experimental therapies for efficacy screening with respect to the same patient population. Usually, the resources for clinical trials are limited, so that we may want to choose only small number of therapies, ideally one therapy, to be compared with a standard therapy through a phase III trial. In this setting, we may take one of two approaches: (i) Conduct multiple separate phase II trials, one for each experimental therapy, and evaluate them independently using a standard phase II trial design method for a single-arm phase II trial; (ii) Conduct a single phase II trial with multiple arms, randomize patients into the arms, and choose the best arm(s) using a selection method. The former approach requires more research resources due to the multiplicity of the studies. Also, the individual phase II trials may potentially have different patient characteristics, and the comparison among different therapies can be biased.

To avoid these issues, the second approach is attractive. However, the statistical approaches for analyzing randomized phase II trials are limited. Simon, Wittes and Ellenberg (1985) consider randomizing n patients to each of K treatment arms through a single stage and picking the winner, the arm with the largest estimated response rate, among them. this approach is based on the statistical methods of ranking and selection, the basic concepts of which were

¹E-mail: E-mail: sinho.jung@duke.edu.

introduced over 50 years ago by Beckhofer (1954), with a substantial literature since that time. They show that, depending on the design setting, $n = 16$ to 70 patients are required for a 0.9 correct selection probability when there exists a difference of 0.15 in response rate among the K arms. Liu, LeBlanc and Desai (1999) point out that this approach has a high selection probability even when the treatment arms have the same response rates. Sargent and Goldberg (2001) consider a similar approach by allowing selection based on other factors when the difference in observed response rates is small.

Thall, Simon and Ellenberg (1989) consider studies with one control and K experimental arms. In the first stage, n_1 patients are randomized to each of K experimental arms, and the winner is chosen for the second stage if its observed efficacy is larger than that for the historical control by 10%. The trial is stopped early if the winner does not satisfy this condition. In the second stage, n_2 patients are randomized to each of the control arm and the selected experimental arm from stage 1, and one-sided testing is conducted to see if the experimental arm is better than the control. They require $n_1 = 30$ to 80 patients and $n_2 = 90$ to 140 patients under different design setting.

Palmer (1991) proposes a two-stage design for selection of the best of three treatments. In stage 1, cohorts of three patients are randomized to Arms A, B and C, and a decision is made to continue to accrue the next cohort or to stop and choose the better two arms. In stage 2, cohorts of two patients are randomized to the two arms chosen at stage 1, and a decision is made to continue to accrue the next cohort or to stop and choose the winner. Given the maximum number of patients available for the study, the stopping time for each stage is chosen to minimize the number of future failures using a Bayesian approach. This method requires rapid determination of responses to be able to apply the sequential tests.

Steinberg and Venzon (2002) propose two-stage designs for a phase II trial with two experimental arms. In stage 1, n_1 patients are randomized to each arm. The trial is stopped after stage 1 if the difference in number of responders between the two arms are larger than d , which is chosen so that, when the two arms have a difference of 0.15 in response rate, the probability of selecting the inferior arm is controlled at a specified level. Otherwise, the trial proceeds to stage 2 to randomize an additional n_2 patients to each arm. After stage 2, the winner is chosen based on the cumulative responses through the two stages. Given $n = n_1 + n_2$, one can choose $n_1 = n_2 = n/2$ or to minimize the expected sample size for the specified response rates with 0.15 of difference. This approach does not control the overall error probabilities through the two stages.

Most of these existing methods do not accurately control the type I error and the power for the whole selection procedure. Furthermore, they do not allow unequal designs among different arms. We propose exact and efficient between-arm comparison methods for analyzing randomized phase II trials designed for independent evaluation of each arm. The proposed methods can be used for comparing the response data from multiple single-arm trials on competing therapies with similar patient populations as well. We use the uniformly minimum variance unbiased estimator (UMVUE) since, as shown by Jung and Kim (2004), for 2-stage phase II trial designs, the maximum likelihood estimator (MLE) can be seriously biased, and the efficiency of UMVUE is comparable to that of MLE. In Section 2, we briefly review the UMVUE for multi-stage designs. We derive between-arm comparison methods under various conditions in Section 3. Some numerical studies are conducted in Section 4.

2 UMVUE - Review

We consider two-stage design for a single-arm phase II trial in this section. For an experimental cancer therapy, let p_0 denote the maximum unacceptable response rate and p_1 denote the minimum acceptable response rate ($p_0 < p_1$). Also, let p denote the true response rate of the

therapy. A typical two-stage phase II trial is conducted as follows. During stage 1, n_1 patients are enrolled and treated. If the number of responders is less than or equal to a_1 , the trial is terminated for lack of efficacy and it is concluded that the treatment does not warrant further investigation (i.e., accept $H_0: p = p_0$). Otherwise, the study is continued to stage 2 during which an additional n_2 patients are enrolled and treated. If the cumulative number of responders after stage 2 does not exceed a , it is concluded that the treatment lacks sufficient efficacy (i.e., accept $H_0: p = p_0$). Otherwise, it is concluded that the treatment has a sufficient efficacy, and the treatment will be considered for further investigation in subsequent trials (i.e., accept $H_1: p = p_1$).

Refer to Simon (1989), Jung, Carey and Kim (2001), Jung et al. (2004) for the search of optimal two-stage designs. One may employ an upper boundary to stop the trial early when a significantly high efficacy is observed from stage 1 (Chang et al, 1987; Spiegelhalter, Freedman, Blackburn, 1986). However, there being no compelling ethical argument and thus rarely used, we consider early stopping only for lack of efficacy in this paper.

A two-stage design is defined by the number of patients to be accrued during stages 1 and 2, n_1 and n_2 , and the boundary values a_1 and a ($a_1 < a$). So, we specify a two-stage design by $(a_1/n_1, a/n)$, where $n = n_1 + n_2$, called the maximum sample size. Let M denote the stopping stage and $S = S_M$ denote the total number of responders accumulated up to the stopping stage.

The UMVUE $\hat{p} = \hat{p}(m, s)$, of p is given as

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

where $u \wedge v = \min(u, v)$ and $u \vee v = \max(u, v)$. Refer to Jung and Kim (2004) for details. The distribution of the UMVUE is derived using the probability mass function of (M, S) in a two-stage design with lower stopping boundaries only is given as

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

3 Comparison and Selection for Randomized Phase II Designs

We consider two-arm randomized phase II studies, each arm with a two-stage design for independent evaluation as the primary objective. Following example study is used throughout this section.

Example 1

Suppose that we randomize non-Hodgkin lymphoma patients who relapsed from a rituximab-containing combination regimen to rituximab alone (Arm R, $n = 90$) or ritux-imab +lenalidomide (Arm R+L, $n = 45$) with 2-to-1 probability. The two arms have the following two-stage designs:

Arm R: $(a_1/n_1, a/n) = (10/57, 19/90)$ for 4% type I error at $p_0 = 0.15$ and 95% power at $p_1 = 0.30$.

Arm R+L: $(a_1/n_1, a/n) = (4/21, 10/45)$ for 5% type I error at $p_0 = 0.15$ and 89% power at $p_1 = 0.35$.

Arm R is a potential control arm for a future phase III trial in case Arm R+L is accepted in this trial, but it is included in this phase II trial because there is not enough historical data on the regimen. Twice as many patients will be accrued to Arm R than to Arm R+L to allow more precise estimation of the clinical parameters to be used in designing a future phase III trial. Arm R+L may not be investigated further if it does not seem to be more efficacious than Arm R. We want to compare the two arms accounting for the two-stage design for each arm.

In general, we call the two arms x and y , respectively. For an outcome (m_k, s_k) , let $\hat{p}_k = \hat{p}_k(m_k, s_k)$ denote the UMVUE for the true response probabilities p_k in arm $k (= x, y)$.

3.1 When Both Arms Have Identical Two-Stage Designs—In this subsection, we assume that the two arms have the same two-stage design $(a_1/n_1, a/n)$ for independent evaluation.

3.1.1 When One Arm Is a Control: One may want to conduct a randomized phase II trial to evaluate an experimental therapy comparing with a prospective control. This may happen when the control therapy has been used as a standard without formal evaluation through a prospective study, or needs more testing in an extended patient population. In this case, the prospective control arm may also be evaluated using a standard two-stage design for phase II trials.

When such a trial is completed, we may want to test if the experimental arm (arm y) is better than the control (arm x) or not. The hypotheses associated with this type of comparison are

$$H_0: p_y \leq p_x \text{ vs. } H_a: p_y > p_x.$$

This is a one-sided test. So in this case, we usually would not want to accept the experimental arm y if it is not accepted in the independent evaluation. Thus, we want to accept the experimental arm (or, reject H_0) if it is accepted in the independent evaluation, i.e. $m_y = 2$ and $s_y > a$, and $\hat{p}_y - \hat{p}_x \geq c$ for a chosen critical value c . Let $\mathcal{D} = \{(m, s): m = 1, 0 \leq s \leq a_1\} \cup \{(m, s): m = 2, a_1 + 1 \leq s \leq n\}$ denote the sample space of each arm defined by the design $(a_1/n_1, a/n)$. Then, given a true response probability $p_x = p_y = p$ under H_0 , the probability of rejecting H_0 is

$$= \sum_{(m_x, s_x) \in \mathcal{D}} \sum_{(m_y, s_y) \in \mathcal{D}} I\{\hat{p}_y(m_y, s_y) - \hat{p}_x(m_x, s_x) \geq c, m_y = 2, s_y > a\} f(m_x, s_x | p) f(m_y, s_y | p), \quad (1)$$

where $I(\cdot)$ is the indicator function and $f(m, s | p)$ denotes the probability mass function of (M, S) under the common two-stage designs. More generally, the probability of an event A in $\mathcal{D}$

² is calculated as

$$P(A|p_x, p_y) = \sum_{(m_x, s_x) \in \mathcal{D}} \sum_{(m_y, s_y) \in \mathcal{D}} I\left\{\left((m_x, s_x), (m_y, s_y)\right) \in A\right\} f(m_x, s_x|p_x) f(m_y, s_y|p_y).$$

In contrast to common asymptotic tests, such as the two-sample t-test, the operating characteristics of our exact test depends on the null response probability p , an unknown nuisance parameter. In order to remove the nuisance parameter, we control the type I error by maximizing the probability in (1) over the whole parameter space $p \in [0, 1]$, or over a subset of interest $\mathcal{Q} \subset [0, 1]$. See Berger and Boos (1994) for the rationale for such an approach. Given α , we want to choose a critical value $c = c_\alpha$ so that the probability of accepting arm y is no larger than α under H_0 , i.e.

$$\begin{aligned} & P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c_\alpha, M_y=2, S_y>a|H_0\} \\ & = \max_{p \in \mathcal{I}} P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c_\alpha, M_y=2, S_y>a|p_x=p_y=p\} \leq \alpha. \end{aligned} \quad (2)$$

We will refer to probability (2) as the type I error. Let p_0 denote the response rate of a historical control. Then, we may choose a small interval such as $\mathcal{Q} = [p_0 - 0.2, p_0 + 0.2]$. In our experience, the maximum type I error usually occurs within this range. Of course, if we want type I error control under any possible situation, we have to choose $\mathcal{Q} = [0, 1]$. We use the latter in this paper.

Let $H(c) = \max_{p \in \mathcal{Q}} h(c|p)$. Obviously, $h(c|p)$ is monotone in c . Given c , however, $h(c|p)$ can have local maxima over $p \in \mathcal{Q}$. For example, when both arms have the same design as that of Arm R+L in Example 1, $(a_1/n_1, a/n) = (4/21, 10/45)$, Figure 1 displays $h(c = 0.1|p)$ over $p \in [0, 1]$. Note that there are two maxima. So, given α , calculation of the critical value c_α requires a 2-stage numerical search procedure. For a given critical value c , $H(c)$ is calculated by the grid search for the maximum of $h(c|p)$ in the range of $p \in [0, 1]$. For any $p \in [0, 1]$, $h(c|p)$ is monotone in c , so that $H(c)$ is also monotone in c . Hence the critical value $c = c_\alpha$ satisfying $H(c_\alpha) = \alpha$ can be obtained by the bisection method.

Given p_x and $p_y = p_x + \Delta$ ($\Delta > 0$), the probability of correct comparison, called the power, is calculated as

$$1 - \beta = P(\widehat{p}_y - \widehat{p}_x \geq c_\alpha, M_y=2, S_y>a|p_x, p_y=p_x+\Delta).$$

Suppose that arm y is accepted in the independent evaluation, and $\hat{c} = \hat{p}_y - \hat{p}_x$ denotes the observed difference from the data. Then, one may want to see how significant the evidence is against H_0 . To this end, we may calculate a p-value by

$$\text{p-value} = \max_{0 \leq p \leq 1} P\{\widehat{p}_y(M_y, S_y) - \widehat{p}_x(M_y, S_y) \geq \hat{c} | p_x = p_y = p\}.$$

Example 2: Suppose that arm x is a control and arm y is an experimental therapy, both with the same two-stage design $(a_1/n_1, a/n) = (4/21, 10/45)$ as in Arm R+L of Example 1. With $\alpha = 0.1$, we have $c_\alpha = 0.1520$ and the Type I error is maximized at $p_x = p_y = 0.2692$. With $\Delta = 0.2$, the power is 0.669 for $(p_x, p_y) = (0.15, 0.35)$, 0.649 for $(p_x, p_y) = (0.2, 0.4)$, and 0.639 for $(p_x, p_y) = (0.25, 0.45)$. With $\Delta = 0.25$, the power is 0.809 for $(p_x, p_y) = (0.15, 0.4)$, 0.796 for $(p_x,$

$p_y) = (0.2, 0.45)$, and 0.800 for $(p_x, p_y) = (0.25, 0.5)$. When we have $(m_x, s_x) = (2, 12)$ ($\hat{p}_x = 0.295$), we have p-value = 0.3064 if $(m_y, s_y) = (2, 15)$ ($\hat{p}_y = 0.342$); p-value = 0.1123 if $(m_y, s_y) = (2, 20)$ ($\hat{p}_y = 0.445$); and p-value = 0.0145 if $(m_y, s_y) = (2, 25)$ ($\hat{p}_y = 0.556$).

Note that the above comparison rule controls the type I error of selecting the experimental arm when both arms have an equal response rate. This rule may be considered too strict. A phase II trial is designed not to show the superiority of an experimental therapy compared to the control, but to screen out ineffective therapies. If an experimental therapy is shown to have no worse efficacy than the control, its superiority may be investigated in a phase III trial using a more definitive endpoint, such as overall survival. In this sense, one may want to loosen the control of type I error somewhat in the phase II design. Let $\delta(> 0)$ denote the maximum clinically insignificant difference in response rate, e.g. $\delta = 0.05$. Suppose that we do not care about falsely accepting arm y as far as p_y is within δ of p_x , i.e. $p_y > p_x - \delta$. In this case, the hypotheses may be modified to

$$H_0: p_y \leq p_x - \delta \text{ vs. } H_a: p_y > p_x - \delta.$$

We choose a critical value $c = c_\alpha$ satisfying

$$\max_{p_x, p_y \in I} P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c_\alpha, M_y=2, S_y > a | p_x, p_y = p_x - \delta\} \leq \alpha.$$

Given p_x and $p_y = p_x + \Delta$, the power is calculated as

$$1 - \beta = P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c_\alpha, M_y=2, S_y > a | p_x, p_y = p_x + \Delta\}.$$

For an observed difference $\hat{c} = \hat{p}_y - \hat{p}_x$, the p-value is calculated as

$$\text{p-value} = \max_{p_x, p_y \in I} P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq \hat{c} | p_x, p_y = p_x - \delta\}.$$

We will allow maximum clinically insignificant difference δ in the remainder of this paper if not stated otherwise.

Example 3: Consider Example 2 with $\delta = 0.05$. With $\delta = 0.05$ and $\alpha = 0.1$, we have $c_\alpha = 0.0925$ and the Type I error is maximized at $(p_x, p_y) = (0.3138, 0.2638)$. With $\Delta = 0.2$, the power is 0.799 for $(p_x, p_y) = (0.15, 0.35)$, 0.820 for $(p_x, p_y) = (0.2, 0.4)$, and 0.827 for $(p_x, p_y) = (0.25, 0.45)$. When we have $(m_x, s_x) = (2, 12)$, we have p-value = 0.1640 if $(m_y, s_y) = (2, 15)$; p-value = 0.0529 if $(m_y, s_y) = (2, 20)$; and p-value = 0.0051 if $(m_y, s_y) = (2, 25)$.

3.1.2 When Both Arms Are Experimental: Suppose now that there are two experimental therapies, x and y, under investigation. The primary objective is to evaluate each therapy compared to a historical control. As a secondary analysis, we want to compare the two experimental arms and choose one that will be investigated further in a phase III trial. Given the maximum clinically negligible difference δ , the hypotheses may be expressed as

$$H_0: |p_x - p_y| \leq \delta \text{ vs. } H_a: |p_x - p_y| > \delta.$$

In this case, the associated testing is two-sided. As in the one-sided case, we do not want to select an experimental arm if it is not accepted in the independent evaluation. That is, we want to select an experimental arm if it is accepted in the independent evaluation and the UMVUE is significantly larger than that of the other arm.

For a chosen critical value c , we select arm x if

$$A_x = \{M_x = 2, S_x > a, \widehat{p}(M_x, S_x) - \widehat{p}(M_y, S_y) \geq c\},$$

is true, and arm y if

$$A_y = \{M_y = 2, S_y > a, \widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c\}$$

is true. Since the two designs are identical for each arm, $(a_1/n_1, a/n)$, the error probabilities $P(A_x | p_x, p_y = p_x + \delta)$ and $P(A_y | p_x, p_y = p_x - \delta)$ are identical. Using this result, we obtain the critical value $c = c_\alpha$ so that the false selection probability under H_0 does not exceed α , i.e.

$$\begin{aligned} & 2P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c, M_y = 2, S_y > a | H_0\} \\ = & 2 \max_{p_x, p_y \in \mathcal{I}} P\{\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c, M_y = 2, S_y > a | p_x, p_y = p_x - \delta\} \leq \alpha. \end{aligned}$$

Note that probabilities $P(A_x)$ and $P(A_y)$ will be unequal if two arms have different designs. Cases with different designs will be discussed in the next section.

Given p_x and $p_y = p_x + \Delta$, the power is calculated as

$$1 - \beta = P(\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq c_\alpha, M_y = 2, S_y > a | p_x, p_y = p_x + \Delta)$$

Suppose that arm y is accepted in the independent evaluation and $\hat{c} = \hat{p}_y - \hat{p}_x (> 0)$ denotes the observed difference in UMVUE from a randomized phase II trial. Then, proceeding as before, we calculate

$$\begin{aligned} \text{p-value} &= 2 \max_{p_x, p_y = p_x - \delta \in \mathcal{I}} P(\widehat{p}(M_y, S_y) - \widehat{p}(M_x, S_x) \geq \hat{c} | p_x, p_y = p_x - \delta), \\ &= 2 \max_{p_x, p_y = p_x - \delta \in \mathcal{I}} \sum_{(m_x, s_x) \in \mathcal{D}(m_y, s_y) \in \mathcal{D}} \sum I\{\widehat{p}_y(m_y, s_y) - \widehat{p}_x(m_x, s_x) \geq \hat{c}\} \\ &\quad \times f(m_x, s_x | p_x) f(m_y, s_y | p_y = p_x - \delta). \end{aligned}$$

We select neither arm if both arms are rejected in the independent evaluation, and select both arms if both arms are accepted in the independent evaluation and $|\hat{p}_y - \hat{p}_x| < c_\alpha$.

Example 4: Consider Example 3, but with both arms considered as experimental. With $\delta = 0.05$ and $\alpha = 0.1$, we have $c_\alpha = 0.1520$ and the Type I error is maximized at $(p_x, p_y) = (0.2565, 0.3065)$, where the order is unimportant. With $\Delta = 0.2$, the power is 0.669 for $(p_x, p_y) = (0.15, 0.35)$, 0.649 for $(p_x, p_y) = (0.20, 0.40)$, and 0.639 for $(p_x, p_y) = (0.25, 0.45)$. When we observe $(m_x, s_x) = (2, 12)$, we have p-value = 0.3280 if $(m_y, s_y) = (2, 15)$; p-value = 0.1058 if $(m_y, s_y) = (2, 20)$; and p-value = 0.0102 if $(m_y, s_y) = (2, 25)$.

If we choose $\delta = 0.1$, then we have $c_\alpha = 0.0925$ for $\alpha = 0.1$, and the Type I error is maximized at $(p_x, p_y) = (0.2460, 0.3460)$, where the order is impertinent. With $\Delta = 0.2$, the power is 0.799 for $(p_x, p_y) = (0.15, 0.35)$, 0.820 for $(p_x, p_y) = (0.20, 0.40)$, and 0.827 for $(p_x, p_y) = (0.25, 0.45)$. When we observe $(m_x, s_x) = (2, 12)$, we have p-value = 0.1501 if $(m_y, s_y) = (2, 15)$, p-value = 0.0442 if $(m_y, s_y) = (2, 20)$, and p-value = 0.0032 if $(m_y, s_y) = (2, 25)$.

3.2 When the Two Arms Have Different Two-Stage Designs—In a randomized phase II trial, we may want to use different designs for different arms. For example, we may want to have more patients in the control arm to allow more efficient estimation of parameters in patient subgroups to be used in designing a phase III trial. Or, we may want to use a less strict early stopping rule in the control arm. If we want to compare two experimental therapies evaluated by separate single-arm phase II trials, it is very likely that the two trials will have different designs. In this section, we consider selection problems when two arms have different 2-stage designs.

In Section 3.1, we considered phase II trials randomizing patients to two arms with exactly the same two-stage designs for independent evaluation. In this case, we did not want to select an arm that is rejected in the independent evaluation. However, when the two arms have different two-stage designs, the selection rules in this section are based only on the comparison of the estimators of the response response rates.

3.2.1 When One Arm Is a Control: As before, let x be the control arm and y the experimental arm and, for a maximal clinically negligible difference δ , we want to test

$$H_0: p_y \leq p_x - \delta \text{ vs. } H_a: p_y > p_x - \delta.$$

We choose a critical value $c = c_\alpha$ satisfying

$$\begin{aligned} & \max_{0 \leq p_x < p_y \leq 1} P(\widehat{p}_y - \widehat{p}_x \geq c_\alpha | p_x = p_y + \delta) \\ = & \max_{0 \leq p_x < p_y \leq 1} \sum_{(m_x, s_x) \in \mathcal{D}_x} \sum_{(m_y, s_y) \in \mathcal{D}_y} I(\widehat{p}_y(m_x, s_x) - \widehat{p}_x(m_y, s_y) \geq c_\alpha) f_x(m_x, s_x | p_y + \delta) f_y(m_y, s_y | p_y) \leq \alpha, \end{aligned}$$

where
 $\mathcal{D}$

k , $\widehat{p}_k(\cdot, \cdot)$, and $f_k(\cdot, \cdot | \cdot)$ are design-specific sample space, UMVUE, and probability mass function, respectively, for arm $k = x, y$.

The power for Δ and p_x ($p_y = p_x + \Delta$) defined as

$$1 - \beta = P(\widehat{p}_y - \widehat{p}_x \geq c_\alpha | p_x, p_y)$$

can be similarly calculated as the type I error. For an observed difference, $\hat{c} = \hat{p}_y - \hat{p}_x$ the p-value is calculated as

$$\text{p-value} = \max_{0 \leq p_x, p_y \leq 1} P(\hat{p}_y - \hat{p}_x \geq \hat{c} | p_x = p_y + \delta)$$

Example 5: Consider $\delta = 0.05$ in Example 1. Then with $\alpha = 0.1$, we have $c_\alpha = 0.0717$ and the Type I error is maximized at $(p_x, p_y) = (0.2185, 0.1685)$. With $\Delta = 0.2$, the power is 0.933 for $(p_x, p_y) = (0.25, 0.45)$, 0.926 for $(p_x, p_y) = (0.30, 0.50)$, and 0.922 for $(p_x, p_y) = (0.35, 0.55)$. Table 1 displays p-values for our exact method.

3.2.2 When Both Arms Are Experimental: Suppose that both arms are experimental with different designs. For a maximal difference clinically negligible δ , we want to test

$$H_0: |p_x - p_y| \leq \delta \text{ vs. } H_a: |p_x - p_y| \geq \delta.$$

We choose a critical value $c = c_\alpha$ satisfying

$$\max_{0 \leq p_x, p_y \leq 1} \{P(\hat{p}_y - \hat{p}_x \geq c_\alpha | p_x = p_y + \delta) + P(\hat{p}_x - \hat{p}_y \geq c_\alpha | p_y = p_x + \delta)\} \leq \alpha. \quad (3)$$

Note that the two misselection errors in the left hand side of (3) are not the same if the two arms have different designs. We fail to select one arm against the other if $|\hat{p}_x - \hat{p}_y| < c_\alpha$.

The power for Δ and p_x ($p_y = p_x + \Delta$),

$$1 - \beta = P(\hat{p}_y - \hat{p}_x \geq c_\alpha | p_x, p_y).$$

For an observed difference, $\hat{c} = |\hat{p}_x - \hat{p}_y|$, the p-value is calculated as

$$\text{p-value} = \max_{0 \leq p_x, p_y \leq 1} \{P(\hat{p}_y - \hat{p}_x \geq \hat{c} | p_x = p_y + \delta) + P(\hat{p}_x - \hat{p}_y \geq \hat{c} | p_y = p_x + \delta)\}.$$

Example 6: Suppose that both arms in Example 1 are experimental. Then with $\delta = 0.05$ and $\alpha = 0.1$, we have $c_\alpha = 0.1174$ and the Type I error is maximized at $(p_x, p_y) = (0.2775, 0.2275)$, where the order is impertinent. With $\Delta = 0.2$, the power is 0.826 for $(p_x, p_y) = (0.25, 0.45)$, 0.831 for $(p_x, p_y) = (0.30, 0.50)$, and 0.838 for $(p_x, p_y) = (0.35, 0.55)$. Table 2 reports p-values for some chosen outcomes for our exact method.

3.3 Extension to More than Two Arms—In this section we consider two situations, one where all arms are experimental and another where one of them is a control. Each arm is independently compared to a historical control through a two-stage design. For arm $k = 0, 1, \dots, K$, let $\mathcal{D}_k$, $\hat{p}_k(\cdot, \cdot)$, and $f_k(\cdot, \cdot)$ denote the sample space, UMVUE, and probability mass function, respectively, which are specific to the design of each arm. If all arms have the same two-stage designs, we can drop the subscripts.

3.3.1 When There Are One Control Arm and K Experimental Arms: Suppose that patients are randomized to a control (Arm 0) and K experimental arms (Arms 1, ..., K). We want to identify experimental arms that are significantly efficacious compared to the control arm. When $K \geq 2$, we have to control the familywise error rate (FWER) to adjust for the multiplicity of the testing. The marginal type I error control applied in the previous sections will increase the misselection probability. For a maximal difference clinically negligible δ , we want to test

$$H_0: p_k \leq p_0 - \delta \text{ for all } k=1, \dots, K$$

against

$$H_a: p_k > p_0 - \delta \text{ for some } k=1, \dots, K.$$

Given a FWER level α , such as 0.1, we accept Arm $k (= 1, \dots, K)$ if $\hat{p}_k - \hat{p}_0 \geq c_\alpha$ where the critical value $c = c_\alpha$ satisfies

$$\begin{aligned} & \max_{0 \leq p_0, p_1, \dots, p_K \leq 1} P(\max_{1 \leq k \leq K} \hat{p}_k - \hat{p}_0 \geq c | p_0, p_1 = \dots = p_K = p_0 + \delta) \\ = & \max_{0 \leq p_0, p_1, \dots, p_K \leq 1} \sum_{(m_0, s_0) \in \mathcal{D}_0} \cdots \sum_{(m_K, s_K) \in \mathcal{D}_K} I(\max_{1 \leq k \leq K} \hat{p}_k(m_k, s_k) - \hat{p}_0(m_0, s_0) \geq c) \prod_{k=0}^K f_k(m_k, s_k | p_k) \leq \alpha. \end{aligned}$$

If more than one arm is accepted, we may conduct pairwise comparisons among accepted arms to identify a smaller number of arms for a phase III trial as described in 3.1.2 and 3.2.2 as a secondary analysis.

Under a specific alternative hypothesis,

$$H_a: p_k = p_0 + \Delta_k \text{ for } k=1, \dots, K$$

with $\Delta_k > 0$, the power is obtained as

$$1 - \beta = \max_{0 \leq p_0, p_1, \dots, p_K \leq 1} P(\max_{1 \leq k \leq K} \hat{p}_k - \hat{p}_0 \geq c_\alpha | p_0, p_1, \dots, p_K).$$

3.3.2 When All K Arms Are Experimental: Suppose that patients are randomized to K experimental arms. In this case, we want to test

$$H_0: p_1 = \dots = p_K$$

against

$$H_a: p_k \neq p_{k'} \text{ for } 1 \leq k \neq k' \leq K.$$

We reject H_0 if $\max_{1 \leq k \leq K} \hat{p}_k - \min_{1 \leq k \leq K} \hat{p}_k \geq c_\alpha$, where critical value $c = c_\alpha$ satisfies

$$\max_{0 \leq p_1 = \dots = p_K \leq 1} P(\max_{1 \leq k \leq K} \hat{p}_k - \min_{1 \leq k \leq K} \hat{p}_k \geq c | p_1 = \dots = p_K) \leq \alpha.$$

Under a specific alternative hypothesis, $H_a: p_1, \dots, p_K$, the power is obtained as

$$1 - \beta = \max_{0 \leq p_1, \dots, p_K \leq 1} P(\max_{1 \leq k \leq K} \hat{p}_k - \hat{p}_0 \geq c_\alpha | p_1, \dots, p_K).$$

4 Discussion

We propose between-arm comparison methods for a randomized phase II trial. If one wants to compare two therapies evaluated through two separate single-arm phase II trials, our methods can be used if the two trials are conducted with similar populations. Each arm to be compared may have a two-stage design for independent evaluation of the therapy, so that statistical procedures based on single-stage designs, such as two-sample t-test, may result in biased results. Our methods accurately compare two arms reflecting the design aspect and the small sample sizes. We have considered two-stage designs, but extension to multi-stage designs is straightforward. The between-arm comparison proposed in this paper is conducted when competing experimental therapies are independently evaluated compared to a historical control. Jung (2007) proposed design methods when patients are randomized between a prospective control and experimental therapies, and each experimental arm is compared with the control through multiple stages.

References

- Beckhofer RE. A single-sample multiple decision procedure for ranking means of normal populations with known variances. *Annals of Mathematical Statistics* 1954;25:16–39.
- Berger R, Boos DD. P values maximized over a confidence set for the nuisance parameter. *J of American Statistical Association* 1994;89:1012–1016.
- Chang MN, Therneau TM, Wieand HS, Cha SS. Designs for group sequential phase II clinical trials. *Biometrics* 1987;43:865–874. [PubMed: 3427171]
- Jung SH. Randomized phase II trials with a prospective control. To appear in *Statistics in Medicine*. 2007
- 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]
- Jung SH, Kim KM. On the estimation of the binomial probability in multistage clinical trials. *Statistics in Medicine* 2004;23:881–896. [PubMed: 15027078]
- 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]
- Liu PY, LeBlanc M, Desai M. False positive rates of randomized phase II designs. *Controlled Clinical Trials* 1999;20:343–352. [PubMed: 10440561]
- Palmer CR. A comparative phase II clinical trials procedure for choosing the best of three treatments. *Statistics in Medicine* 1991;10:1327–1340. [PubMed: 1925164]
- Sargent DJ, Goldberg RM. A flexible design for multiple armed screening trials. *Statistics in Medicine* 2001;20:1051–1060. [PubMed: 11276035]
- Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989;10:1–10. [PubMed: 2702835]
- Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treatment Reports* 1985;69:1375–1381. [PubMed: 4075313]

- Spiegelhalter DJ, Freedman LS, Blackburn PR. Monitoring clinical trials: Conditional or Predictive power? *Controlled Clinical Trials* 1986;7:8–17. [PubMed: 3956212]
- Steinberg SM, Venzon DJ. Early selection in a randomized phase II clinical trial. *Statistics in Medicine* 2002;21:1711–1726. [PubMed: 12111907]
- 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. [PubMed: 2765637]

Figure 1.
Plot of $h(c|p)$ for $c = 0.1$ and p between 0 and 1

Table 1

p-values for some chosen outcomes for comparing Arm x (control) with $(a_1/n_1, a/n)$ with $(10/57, 19/90)$ and Arm y (experimental) with $(a_1/n_1, a/n) = (4/21, 10/45)$ at $\alpha = 0.1$ and $\delta = .05$.

Arm x			Arm y			p-value
m	s	$\hat{p}(m, s)$	m	s	$\hat{p}(m, s)$	
2	20	0.230	2	11	0.283	0.1413
2	20	0.230	2	12	0.295	0.1169
2	20	0.230	2	13	0.309	0.0904
2	20	0.230	2	14	0.325	0.0675
2	20	0.230	2	15	0.342	0.0447
2	20	0.230	2	20	0.445	0.0017

Table 2
p-values for some chosen outcomes for comparing two experimental arms with two-stage designs (a_1/n_1 , a/n) = (10/57, 19/90) and (a_1/n_1 , a/n) = (4/21, 10/45) at $\alpha = 0.1$ and $\delta = .05$.

<i>m</i>	Arm <i>x</i>		Arm <i>y</i>		p-value
	<i>s</i>	$\hat{p}(m, s)$	<i>m</i>	<i>s</i>	
2	33	0.333	2	20	0.2180
2	33	0.333	2	25	0.0277
2	33	0.333	2	30	0.0019
2	33	0.333	2	35	0.0001
2	33	0.333	2	40	0.0000