Received: Added at production | Revised: Added at production | Accepted: Added at production

DOI: xxx/xxxx

ORIGINAL ARTICLE

An improved Bayesian pick-the-winner (IBPW) design for randomized phase II clinical trials

Wanni Lei[1] | Maosen Peng[1,2] | Xi Kathy Zhou*[1]

[1]Department of Population Health Sciences, Weill Cornell Medicine, New York, USA
[2]Department of Biostatistics and Data Science,University of Texas School of Public Health, Houston, TX, USA

Correspondence
*Xi Kathy Zhou, Division of Biostatistics, Department of Population Health Sciences, Weill Cornell Medicine, Email: kaz2004@med.cornell.edu

Abstract

Phase II clinical trials play a pivotal role in drug development by screening a large number of drug candidates to identify those with promising preliminary efficacy for phase III testing. Trial designs that enable efficient decision-making with small sample sizes and early futility stopping while controlling for type I and II errors in hypothesis testing, such as Simon's two-stage design, are preferred. Randomized multi-arm trials are increasingly used in phase II settings to overcome the limitations associated with using historical controls as the reference. However, how to effectively balance efficiency and accurate decision-making continues to be an important research topic. A notable development in phase II randomized design methodology is the Bayesian pick-the-winner (BPW) design that extends a Simon's two-stage based multi-arm design with a Bayesian winner-selection strategy. Despite multiple appealing features, this method cannot easily control for overall type I and II errors for winner selection. Here, we introduce an improved randomized two-stage Bayesian pick-the-winner (IBPW) design that formalizes the winner-selection based hypothesis testing, optimizes sample sizes and decision cut-offs by strictly controlling the type I and type II errors under a set of flexible hypotheses for winner-selection across two treatment arms. Simulation studies demonstrate that our new design offers improved operating characteristics for winner selection while retaining the desirable features of the BPW design.

KEYWORDS:
Bayesian statistics; phase II clinical trial designs; pick-the-winner design; Optimal/Minimax design, Simon's two-stage design

1 | INTRODUCTION

Phase II clinical trials are intermediate-stage clinical studies of drugs at the recommended doses that have passed initial safety evaluation. These studies focus on preliminary efficacy and further safety assessments which are crucial for deciding whether to advance to larger, confirmatory phase III trials. In oncology drug development, it is estimated that 57.6% of drug candidates examined in phase I trials were further examined in phase II, 32.7% of drug candidates evaluated in phase II were tested in phase III while about 35.5% of those examined in phase III won regulatory approval[1]. Considering the vast number of drug candidates and combinations advancing to phase II, yet with a relatively small proportion achieving ultimate success, phase II trial designs critically balance efficiency – favoring a small sample size to minimize cost and time – with the need for robust

preliminary efficacy assessments to identify promising phase III candidates. Moreover, designs that incorporate features such as the possibility of early futility stopping to reduce patients' exposure to ineffective treatment, ease of trial implementation and conduct, and easily accessible decision-making rules, can positively affect their practical adoption.

When designing a phase II oncology trial, important initial considerations include selecting the primary endpoints, the treatments to be evaluated or compared, and the number of interim analyses or stages for decision-making. Prioritizing efficiency, binary endpoints, such as the objective response rate, based on tumor shrinkage[2,3], or the short-term progression-free survival[4], which can be measured reasonably quickly following treatment and may predict clinical benefit, are commonly used. Single-arm designs with historical controls are widely adopted[5] as both type I and II errors of hypothesis tests in such trials can be controlled for with a small sample size. Two-stage designs, allowing early futility termination without compromising trial efficiency, are preferred over single-stage designs.

Two widely used single-arm two-stage phase II trial designs in oncology were proposed by Gehan[6] in 1961 and Simon[7] in 1989. In Gehan's two-stage design, the first stage sample size is determined by controlling the type II error rate in rejecting the treatment with a very low rejection cut-off (typically 0 response) for an expected or desired treatment effect. The second stage sample size is determined based on the desired accuracy in estimating the treatment effect. This design allows early futility stopping with a small sample size in the first stage. Simon's two-stage design determines the optimal sample sizes in each stage and rejection cut-offs by controlling both the type I and type II errors when comparing the treatment against historical controls. This design is noted for its ease of implementation, good operational characteristics with minimized patient numbers under the null hypothesis, and the ability to halt early for futility. As a result, Simon's two-stage design has been a preferred design for phase II oncology trials in the past two decades, as evidenced by its adoption in around 40% of the 347 phase II trials published in 2005, 2010, and 2014 in leading oncology journals[8].

One major limitation of the single-arm design for phase II trials is that the efficacy of the new treatment is compared with a specific target value based on historical studies or investigators' prior experiences. This approach ignores the uncertainties in the target value[9], possible variations in response assessment criteria[10], and potential differences in patient populations. To improve on this, randomized multi-arm designs have been increasingly used in phase II oncology trials in recent years. Lee and Feng's[11] review of 266 randomized phase II oncology trials published from 1986 to 2002 revealed an increase in the number of such trials from 2 in 1986 to 34 in 2002. Ivanova et al.[8] reviewed 56 phase II oncology trials conducted in 2014, and found that 41% of them were randomized clinical trials. Randomization enables potentially more valid comparisons, yet the challenge remains in conducting efficient studies with small sample sizes that confidently differentiate between arms. One strategy is to focus on the alternative hypotheses only. The pick-the-winner design by Simon, Wittes, and Ellenberg (SWE)[10] is based on such an idea and selects one or multiple winners among several treatments based on the ranking and selection theory. The sample size is determined to achieve a high probability of identifying a treatment as the winner if it is superior to all others by a prespecified amount. This design is very efficient and controls the type II error. To enable early futility stopping, Steinberg and Venzon (SV)[12] proposed a new approach based on a two-stage statistical selection theory, allowing for early termination of the less effective arm when response proportions in the two arms are sufficiently different. However, neither SWE nor SV designs are capable of simultaneously controlling both type I and type II error rates. Another strategy is to design the randomized study not for direct between-arm compassions but having each arm follow a single-arm design, often Simon's two-stage design, where efficacy is compared to historical controls. This approach is very widely used in phase II oncology multi-arm studies[8]. An additional strategy as proposed by Zhou et al.[13] in their two-arm BOP2 design is to focus on the finding of decision rules that are optimized for power in carrying out between-arm comparisons based on user-specified sample sizes, number of interim analyses, and prespecified type I error rate.

Recently, Chen et al.[14] introduced a Bayesian pick-the winner (BPW) design for randomized phase II clinical trials that improved on Simon's two-stage single-arm design based multi-arm approach by enabling winner-picking at the end of the trial. A winner is declared if one arm passes the second stage of the trial while the other arm fails either stage 1 or 2, or the posterior probability of the winning arm having a greater response rate passes a predefined threshold. This strategy retains the attractive operating characteristics of Simon's two-stage design for comparing treatments against historical controls, while also enabling the between-arm winner identification at the end of the trial and allowing for the evaluation of winner-picking probabilities under different hypotheses at the design stage.

In practice, for two-arm phase II oncology trials, it can be desirable to find a design that 1) tightly controls the type I and type II errors for winner-picking, rather than for comparisons between treatments and historical controls; 2) adheres to certain optimality criteria in terms of sample sizes and decision rules. Achieving these objectives with existing approaches can be challenging. Firstly, the hypotheses specification is commonly framed in terms of treatment versus historical control[14] or the

arm A versus arm B for the traditional two-arm comparative study, not in terms of winner-picking, i.e., no winner versus having a winner. Secondly, a design optimal for single-arm treatment versus control or arm A versus arm B comparisons may not be optimal for winner-picking. To address these challenges, we have developed an improved two-stage Bayesian pick-the-winner design. This design adopts a flexible approach to hypotheses specification focused on winner-picking and optimizes sample sizes and decision cut-offs while tightly controlling the type I and II errors for winner identification.

The remainder of this article is structured as follows. In the Method section, we detail the flexible hypotheses specifications for winner picking and introduce the algorithms of our improved randomized two-stage Bayesian pick-the-winner design for identifying optimal design parameters. The Simulation section presents comprehensive simulations conducted to assess the performance of our novel design and compare its operating characteristics with those of the designs by Chen et al.[14]. The article concludes with a discussion.

2 | METHOD

Our objective is to determine the design parameters for a two-arm, two-stage randomized phase II oncology trial that meets optimality criteria similar to Simon's two-stage design, i.e., minimal expected or maximal sample sizes, while strictly controlling type I and II errors under hypotheses specified for winner picking between two treatments, rather than conventional direct treatment versus control or treatment A versus treatment B comparisons. We assume that the primary endpoint is binary, for example, the response rates of the study treatments, and the trial is operated similarly to the original BPW design[14] for selecting the winner. Our design differs from existing randomized two-stage designs in two main aspects: first, the hypotheses specification is more flexible, focusing on treatment effect-based winner picking instead of direct treatment effect comparisons; second, our design optimizes the sample sizes and determines the decision cut-offs for each stage by strictly controlling the type I and II errors for correct winner identification. In the following sub-sections, we elaborate on the important methodological components of our design.

2.1 | Hypotheses for winner selection

We consider the setting where treatment B is presumed to be better than treatment A concerning the binary primary endpoint, such as the response rate. We are interested in examining the following hypotheses:

$$H_0 : \text{B is not a winner in comparison with A}, \quad H_1 : \text{B is a winner in comparison with A}.$$

More specifically, we denote P_A and P_B as the response rates of treatments A and B, respectively. Let p_{A_0} and p_{B_0} represent the upper limits of the response rates in arms A and B respectively under the null hypothesis, and p_{A_1} and p_{B_1} denote the lower and upper limits of these rates under the alternative hypothesis, respectively. The hypotheses can be formalized as:

$$H_0 : P_B \le p_{B_0} \quad v.s. \quad P_A \le p_{A_0}, \quad H_1 : P_B > p_{B_1} \quad v.s. \quad P_A \le p_{A_1} \quad ,$$

where the response rate thresholds satisfying $p_{A_0} \le p_{B_0} \le p_{A_1} \le p_{B_1}$ are considered valid. The threshold values can be determined by considering the response rate observed in historical controls or the low rates that should lead to the rejection of the treatments, and the high rate that warrants worth further exploration. We show below through examples that this hypotheses specification framework generalizes the traditional hypotheses for two group comparison and enables the examination of clinically relevant additional scenarios for winner selection through individual specification of the threshold values.

In traditional two group comparisons, the hypotheses being tested are: $P_B \le P_A$ (H_0) and $P_B > P_A$ (H_1), or specifically, $P_B \le p_A$ (H_0) and $P_B \ge p_B$ (H_1), where p_A and p_B denote the threshold values for P_A and P_B, respectively, with $p_B > p_A$. These hypotheses can be written under our aforementioned framework for winner selection by assuming $p_{A_0} = p_{B_0} = p_{A_1} = p_A$ and $p_{B_1} = p_B$. An example of these hypotheses is illustrated below.

$$H_0 : p_{B_0} = 0.05 \quad vs. \quad p_{A_0} = 0.05, \quad H_1 : p_{B_1} = 0.4 \quad vs. \quad p_{A_1} = 0.05 \, . \qquad (HT_a)$$

In practice, when B is an experimental treatment and A is an experimental or control treatment with the response rate not entirely clear for the study population, there are scenarios where hypotheses as specified in HT_a may not accurately capture the trial intention. In one such scenario, clinicians may want to reject both A and B if B is only slightly better than A, where A is assumed to have a similar response rate as the standard of care (SOC). The clinician will consider B as a winner when there

is evidence that B is much better than A. Hypotheses underlying this scenario can be accurately specified within our proposed winner selection framework as posed to the traditional two-group comparison hypotheses framework because three threshold values are involved. An example of the thresholds specified for such hypotheses is shown below:

$$H_0 : p_{B_0} = 0.1 \quad vs. \quad p_{A_0} = 0.05; \quad H_1 : p_{B_1} = 0.4 \quad vs. \quad p_{A_1} = 0.05\,. \tag{HT_b}$$

In another scenario, the clinician may want to reject both B and A if they perform similarly as the standard of care. The clinician also thinks that A may have moderate improvements over SOC while B may have significant improvement over SOC. If the latter is true, B will be considered as a winner. Such study objectives can be translated to the winner selection hypotheses as follows:

$$H_0 : p_{B_0} = 0.1 \quad vs. \quad p_{A_0} = 0.1; \quad H_1 : p_{B_1} = 0.4 \quad vs. \quad p_{A_1} = 0.2\,. \tag{HT_c}$$

It's also possible that clinicians may want to reject both A and B if B is only slightly better than A, but consider B as a winner when A has moderate improvements over SOC and B has significant improvement over SOC, i.e., a combination of HT_b and HT_c with hypotheses as follows:

$$H_0 : p_{B_0} = 0.15 \quad vs. \quad p_{A_0} = 0.1; \quad H_1 : p_{B_1} = 0.4 \quad vs. \quad p_{A_1} = 0.2\,. \tag{HT_{bc}}$$

We show in the next sub-section that these hypotheses enables straight forward calculation of type I and type II errors for winner selection for a winner selection strategy, such as the BPW approach[14].

2.2 | Errors of winner selection

We adopt the same winner selection strategy as the BPW design[14]. The trial operates as follows. In stage 1, a total of $2 \times n_1$ patients are randomized at a 1:1 ratio to arms A and B. If a treatment arm records $\leq r_1$ responses among n_1 patients, it is discontinued for futility. Otherwise, it continues, enrolling additional patients to reach a total sample size of n, through continued 1:1 randomization if both arms advance past stage 1, or direct enrollment if the other arm is halted. At the end of stage 2, if $\leq r$ responses are observed among a total of n patients, the treatment is deemed inadequate and rejected. Otherwise, it is considered as a potential winner. A treatment is declared as a winner if 1) it passes stage 2 while the other treatment fails at either stage 1 or stage 2; or 2) it show superior response rate with the Bayesian posterior probability of a greater response rate comparing to the other treatment exceeding a prespecified threshold δ (e.g., 0.8).

For this winner selection strategy, we can calculate the probabilities of arm B being incorrectly declared as the winner under the null hypothesis (type I error) and correctly declared under the alternative hypothesis (power/1 - type II error), which can be written as the following.

$$\text{power for winner selection} = \Pr(\text{ B wins} \mid H_1)$$
$$\text{type I error for winner selection} = \Pr(\text{ B wins} \mid H_0)$$

Generally, let Y denote the number of responses observed among n patients treated with a treatment having a response rate of p. We can model Y as following a binomial distribution, i.e., $Y \sim Binomial(n, p)$. Given a set of threshold values for the response rates specified under the null and alternative hypotheses, we can determine the probability of arm B being declared as the winner for a design with parameters n, n_1, r, and r_1, as follows.

$$\Pr(\text{ B wins} \mid p_A, p_B, n, n_1, r, r_1) = \text{I} \times \text{II} + \text{III}$$

where each of the terms I, II, and III represent the probability of arm B passing stage 2, the probability of arm A failing either stage 1 or stage 2 and the probability of B being declared as a winner when both arms pass stage 2, respectively. These terms can be calculated as shown below.

$$\text{I} = 1 - \Pr(Y \leq r_1 | n_1, p_B) - \sum_{y=r_1+1}^{min(n_1,r)} \Pr(Y = y | n_1, p_B) \cdot \Pr(Y \leq r - y | n - n_1, p_B)$$

$$\text{II} = \Pr(Y \leq r_1 | n_1, p_A) + \sum_{y=r_1+1}^{min(n_1,r)} \Pr(Y = y | n_1, p_A) \cdot \Pr(Y \leq r - y | n - n_1, p_A)$$

$$\text{III} = \sum_{y_A=r+1}^{n} \sum_{y_B=r+1}^{n} \Pr(Y = y_B | n, p_B) \cdot \Pr(Y = y_A | n, p_A) \cdot I(\Pr(P_B > P_A | y_A, y_B, n) > \delta)\,,$$

Notice that for term III, $\Pr(Y = y_B|n, p_B)$ with $y_B > r$ can be calculated as $\sum_{y=r_1+1}^{min(n_1,y_B)} \Pr(y|n_1, p_B) \cdot \Pr(y_B - y|n - n_1, p_B)$ where y is the stage 1 number of responses in this treatment arm, $\Pr(Y = y_A|n, p_A)$ with $y_A > r$ can be calculated similarly, and $I(\Pr(P_B > P_A|y_A, y_B, n) > \delta)$ is an indicator function which has a value of 1 when the posterior probability of the response rate of treatment B being greater than the response rate of treatment A given data exceeds δ and a value of 0 otherwise.

Assuming the prior probability of the response rate, P, follows a Beta distribution, i.e., Beta(a_0, b_0), the posterior probability of P can be written as the following for the two-stage design:

$$\Pr(P|y, n, n_1, r_1) \propto P^{y+a_0-1}(1-P)^{n_1-y+b_0-1} \cdot C_{n_1}^{y} \cdot I(y \le r_1) + P^{y+a_0-1}(1-P)^{n-y+b_0-1} \cdot \sum_{y_1=r_1+1}^{min(y,n_1)} C_{n_1}^{y_1} C_{n-n_1}^{y-y_1} \cdot I(y > r_1) .$$

Hence, when $y_A > r > r_1$, the posterior probability of P_A follows a Beta distribution, i.e., $P_A|data \sim Beta(y_A + a_0, n - y_A + b_0)$. Similarly, $P_B|data \sim Beta(y_B + a_0, n - y_B + b_0)$ when $y_B > r > r_1$. The posterior probability of B having a higher response rate than A given data, i.e., $\Pr(P_B > P_A|data)$, can be calculated through 1) simulations by sampling P_A and P_B separately from their respective posterior probabilities and calculating the proportion of iterations with $P_B > P_A$; or 2) the following numerical integration $\int_0^1 [1 - \Pr(P_B \le x|n, y_B, a_0, b_0)] \cdot \Pr(P_A = x|n, y_A, a_0, b_0)\mathrm{d}x$. A closed-form solution for $\Pr(P_B > P_A|data)$ is available through formula by Pham-Gia and Turkkan[15] although it involves the evaluations of a complicated two-dimensional generalized hypergeometric function. When preliminary data are lacking, a non-informative prior such as the uniform prior, i.e., Beta(1,1), can be a reasonable choice as the estimates derived with such a prior would typically demonstrate comparable properties as the matching frequentist approach[16]. For the BPW design by Chen et.al.[14], $Beta(1, 1)$ is the default choice.

2.3 | Design parameter optimization

For our randomized two-stage BPW design, similar to Simon's two-stage design, we consider two approaches to optimizing the design parameters (n, n_1, r, r_1), given the response rate thresholds under the null and alternative hypotheses, the constraints in terms of type I and II errors in winner selection, and the Bayesian posterior probability threshold δ, by minimizing either the expected total sample size under the null hypothesis (optimal design) or the maximum total sample size (minimax design). Specifically, the expected total sample size under H_0, denoted as $EN(p_0)$, can be calculated as follows:

$$\begin{aligned} EN(p_0) = & \text{P(Both arms fail in stage 1 | } H_0 \text{)} \times 2n_1 + \\ & \text{P(One arm fails in stage 1 | } H_0) \times (n_1 + n) + \\ & \text{P(Both arms pass stage 1 | } H_0) \times 2n \\ = & P(Y \le r_1|n_1, p_{B_0}) \times P(Y \le r_1|n_1, p_{A_0}) \times 2n_1 + \\ & [P(Y > r_1|n_1, p_{A_0}) \cdot P(Y \le r_1|n_1, p_{B_0}) + P(Y > r_1|n_1, p_{B_0}) \cdot P(Y \le r_1|n_1, p_{A_0})] \times (n_1 + n) + \\ & P(Y > r_1|n_1, p_{A_0}) \times P(Y > r_1|n_1, p_{B_0}) \times 2n . \end{aligned}$$

The optimal design parameters (n, n_1, r_1, r) can be identified through enumeration over a set of feasible designs with $n \in [6, n_{\max}]$, $n_1 \in [3, n_{\max} - 3]$, $r_1 \in [0, n_1 - 1]$, and $r \in [1, n_2 + r_1 - 1]$, where $n_{\max}$ is set to 100 by default in our algorithm. For each feasible design, we calculate the type I and II errors associated with winner selection and the expected total sample size under the null hypothesis. The optimal, or minimax, design is then identified as the one with design parameters that best meet the specified optimality criterion while adhering closely to the constraints regarding type I and II errors in winner selection.

2.4 | Software

We have developed a user-friendly R Shiny application that enables users to readily identify the optimal or minimax design using our improved Bayesian pick-the-winner design method. Additionally, this application allows users to explore the operating characteristics of their customized designs. Both the R functions and the Shiny application are available for download at the following link: https://doi.org/10.5281/zenodo.10047803.

3 | SIMULATION

We carried out an extensive simulation study to evaluate the operating characteristics of our IBPW design in comparison with the BPW design[14] and for testing flexible sets of clinically relevant hypotheses related to winner selection. We considered a range of hypotheses for winner selection based on a binary endpoint, e.g., the response rate. The type I error and type II error rates were set to 10%, and 20% respectively. A $beta(1, 1)$ prior was used to calculate the posterior probability $\Pr(P_B > P_A|data)$ when both treatment arms pass the second stage. The Bayesian posterior probability threshold δ is set to 0.8.

3.1 | IBPW designs strictly control type I and II errors in winner selection

The major difference between the BPW[14] design and our IBPW design lies in the hypotheses being examined and the type I and II error constraints used in determining the optimal design parameters. The BPW design is a straightforward multi-arm adaption of Simon's two-stage single-arm design where the hypotheses center on treatment versus the historical control comparison and the type I and II errors are calculated as the probability of the response rate being greater than a rate specified in the null hypothesis. Our IBPW design, on the other hand, examines hypotheses concerning winner identification between two treatment arms. The type I error and power are calculated as the probability of correctly identifying the true winner arm. We compared the optimal and minimax designs generated by the BPW[14] design method with those generated by our IBPW design method in eight different hypotheses testing scenarios. These scenarios involved type HT_a hypotheses as described in section 2.1, i.e. hypotheses in line with the traditional two-group comparison.

Table 1 presents the design parameters (n, n_1, r, r_1), the expected total sample size for both arms A and B under H_0 $(EN(p_0))$, and the power and type I error (alpha) in winner selection for the optimal and minimax designs identified using the BPW and IBPW approaches. The difference between response rate thresholds of two arms under H_1 is 35% for Scenario 1.1, 30% for Scenarios 1.2 to 1.4, and 20% for Scenarios 1.5 to 1.8.

In general, the designs identified with the BPW and IPBW approaches require substantially smaller total expected or maximum sample size than that required with a traditional single-stage design for two group comparison design at the same error levels albeit the errors have different meanings. For example, for Scenario 1.1, using a traditional single-stage design for two group comparison, a sample size of 14 is required for each arm to achieve 80% power at the one-sided significance level of 0.10 while the BPW and IPBW required a maximum sample size of 8 for each arm.

Additionally, the optimal and minimax designs identified with our IBPW approach strictly control the type I and II errors, whereas the BPW[14] designs are underpowered for winner detection in all scenarios except for Scenario 1.1. The increase in power is achieved with a small increase in either $EN(p_0)$ for the optimal designs or the total maximum sample size for the minimax designs.

More specifically, for the optimal-IBPW designs, the increase in expected total sample size ranges from 0 in Scenario 1.1 to around 7.722 (3.9 per treatment arm) in Scenario 1.7 and the increase in power can be up to 7.29% as for Scenario 1.7 in our simulated scenarios. Interestingly, the increase in $EN(p_0)$ does not necessarily result in an increase in the total maximum sample size. For both Scenarios 1.2 and 1.8, the total maximum sample sizes in the optimal-IBPW designs are smaller than those in the optimal-BPW designs.

For the minimax-IBPW designs, the increase in maximum total sample size ranges from 0 in Scenario 1.1 to 14 (7 per treatment arm) in Scenario 1.6 with an increase in $EN(p_0)$ ranges from -1.25 in Scenario 1.2 to 7.282 in Scenario 1.7 and an increase in power ranges from 0 (Scenario 1.1) to 8.08% (Scenario 1.6).

It should be noted that while it is possible to adjust the power and type I error level of Simon's two-stage design to identify designs that achieve the desired power and type I error for winner selection using the BPW approach, this process is not straightforward, and the designs may not be optimal for winner selection.

3.2 | IBPW identifies optimal designs for flexible winner selection hypotheses

As described in section 2.1, traditional two-group comparison hypotheses (HT_a) may not adequately address the range of scenarios clinicians may want to examine in practice. Specifically, the scenario where a small improvement in arm B over arm A is deemed uninteresting (HT_b), the scenario where it is of interest to correctly identify B as the winner when arm B has a large improvement and arm A has a small improvement (HT_c), and the combination of these two (HT_{bc}) cannot be examined under the traditional hypotheses testing framework. We show that our IBPW approach, employing a flexible hypotheses specification

framework for winner selection, can identify optimal and minimax designs for testing various sets of flexible hypotheses while strictly controlling type I and II errors for winner selection.

Table 2 lists the optimal and minimax designs for a total of 8 scenarios, where Scenarios 2.1 and 2.5, Scenarios 2.2 and 2.6, Scenarios 2.3 and 2.7, and Scenarios 2.4 and 2.8 represent examples of hypotheses types HT_a, HT_b, HT_c, and HT_{bc}, respectively. To assess the sensitivity of designs identified for hypotheses with small differences, response rate threshold differences in Scenarios 2.1-4 (and in Scenarios 2.5-8) are set at 0.05 and limited to 1 or at most 2 parameters. The target type I and II errors for winner selection are set at 0.1 and 0.2, respectively. The table also includes the expected total sample size under H_0, $EN(p_0)$, as well as the actual power and type I error rates for correct winner selection.

All the optimal and minimax designs for the testing scenarios have actual power and type I error rate meet our target. The designs identified are sensitive to small differences in the hypotheses. For example, compared to Scenario 2.1, the optimal and minimax designs for Scenario 2.2 require slightly larger sample sizes, with an increase in $EN(p_0)$ close to a total of 5 subjects and an increase in total maximum sample size of 10, and different stage 1 sample sizes and decision cut-offs due to a small increase in p_{B_0} from 0.1 to 0.15 to capture a clinically uninteresting improvement in arm B while holding other thresholds the same. A similar magnitude of change in p_{A_1} as shown in Scenario 2.3 in comparison with Scenario 2.1 led to optimal and minimax designs with an increase in $EN(p_0)$ by about 5 subjects and total maximum sample size by about 8 subjects, and different stage 1 sample sizes and decisions cut-offs. As expected, when both p_{B_0} and p_{A_1} increases as illustrated in Scenario 2.4 compared to Scenario 2.1, a greater increase in sample sizes is required for the optimal and minimax designs. Similar trends are observed for Scenarios 2.5-8.

The sensitivity of these designs to small difference in hypotheses underscores the importance of a flexible hypotheses specification framework to capture subtle differences in hypotheses and a corresponding tailored approach to identifying designs that meet error control and optimality criteria.

3.3 | Operating characteristics

Table 3 summarizes the operating characteristics of optimal and minimax designs obtained by the IBPW approach under the same eight scenarios in Table 2. Columns 6-11 detail the probabilities of both arms stopping at stage 1 resulting in a total sample size of $2n_1$, only one arm stopping at stage 1 resulting in a total sample size of $n + n_1$, and both arms passing stage 1 resulting in a total sample size of $2n$ under null and alternative hypotheses, respectively. The final four columns show the probabilities of a single arm advancing past stage 2 under a specific hypothesis threshold.

Both the optimal and minimax IBPW designs exhibit desirable operating characteristics. The probabilities of at least one arm stopping early are high under the null hypothesis while the probabilities of at least one arm passing stage 1 are high under the alternative hypothesis. Specifically, the optimal designs generated in our simulation study have an early stopping probability of at least 88.2% in Scenario 2.1 and up to 96.5% in Scenario 2.3 under the null hypothesis. The minimax designs ensure a minimum early stopping probability of 67.5% in Scenario 2.3 and a maximum of 99.4% in Scenario 2.6. In all the IBPW designs, the nonpromising arm has a much higher probability of being screened out in the first stage (see details in Supplementary Tables S1 and S2), thus reducing patient exposure to ineffective treatment.

Within each treatment arm for comparisons against a fixed response rate, the IBPW designs also show favorable operating characteristics. In all scenarios, the probability of each arm passing stage 2 under the null hypothesis generally remains below 10%, except in the optimal design for Scenario 2.1 where it slightly exceeds this at 11.2%. This suggests effective control of single-arm type I error rate for the comparison against a fixed response rate. Given the alternative threshold values p_{B_1} specified in the study scenarios, arm B has a probability of at least 81.3% of passing stage 2 in the optimal and minimax designs, indicating effective control of type II error rate for the comparisons against a fixed response rate.

Uncertainty in hypothesis thresholds can impact the operating characteristics of a design, as demonstrated with the optimal and minimax designs for Scenarios 2.4 and 2.8, detailed in Table 4. With the IBPW optimal design for Scenario 2.4, decreasing p_{B_0} by 0.05 while keeping p_{A_0}, p_{A_1}, and p_{B_1} the same resulted in a reduction in the type I error (from 0.06 to 0.01) for winner selection and the expected total sample size under the null hypothesis (from 24.882 to 22.953) without affecting the power for winner selection. The IBPW designs for Scenario 2.4 have increased power for both the optimal (from 0.813 to 0.839) and minimax (from 0.802 to 0.842) designs when p_{A_1} is assumed to be 0.05 smaller while p_{A_0}, p_{B_0}, and p_{B_1} remain unchanged due to the larger difference in response rates in the alternative. The type I error and $EN(p_0)$ were not affected given the same null hypothesis. The optimal and minimax designs in Scenario 2.8 exhibit similar patterns of change in operating characteristics when similar adjustments were made to the hypothesis thresholds.

3.4 | Effect of delta

We conducted a sensitivity analysis to examine how variations in the delta value affect the design parameters of our optimal-IBPW and minimax-IBPW designs. Figure 1 displays the variations in design parameters (n, n_1, r, r_1) for both designs across the same eight scenarios detailed in Table 2, with delta values ranging from 0.65 to 0.95. The first and second rows of the figure present the results for the minimax and optimal designs, respectively, for Scenarios 2.1 to 2.4, while the third and fourth rows illustrate the designs for Scenarios 2.5 to 2.8.

In most scenarios, a higher delta value generally corresponds to an increase in either n or n_1 for both designs. However, this pattern is not consistent due to the discrete nature of the design parameters and the interplay between sample sizes and response rate cutoffs at each stage of the two-stage design. For the minimax design, although an increase in delta leads to an increase in n or n_1 in most scenarios, there are exceptions. For instance, the design parameters remain unchanged when the value of delta changes from 0.65 to 0.95 in Scenario 2.2, consistent with the very low probability of arm A passing stage 2. An increase in delta from 0.75 to 0.85 resulted in a decrease in n_1 from 16 to 11 and an increase in r from 11 to 12 in Scenario 2.6, and a decrease in both n_1 and r_1 by 1 in Scenario 2.8. Similar patterns of change are observed for the optimal designs. While an increase in delta generally results in an increase in either n or n_1 in most scenarios, interestingly, an increase in delta from 0.85 to 0.95 in Scenario 2.5 led to a decrease in both n_1 and r_1 by 1, and increasing delta from 0.65 to 0.75 in Scenario 2.7 and from 0.75 to 0.85 in Scenario 2.8 resulted in decreases in n, n_1, and r_1.

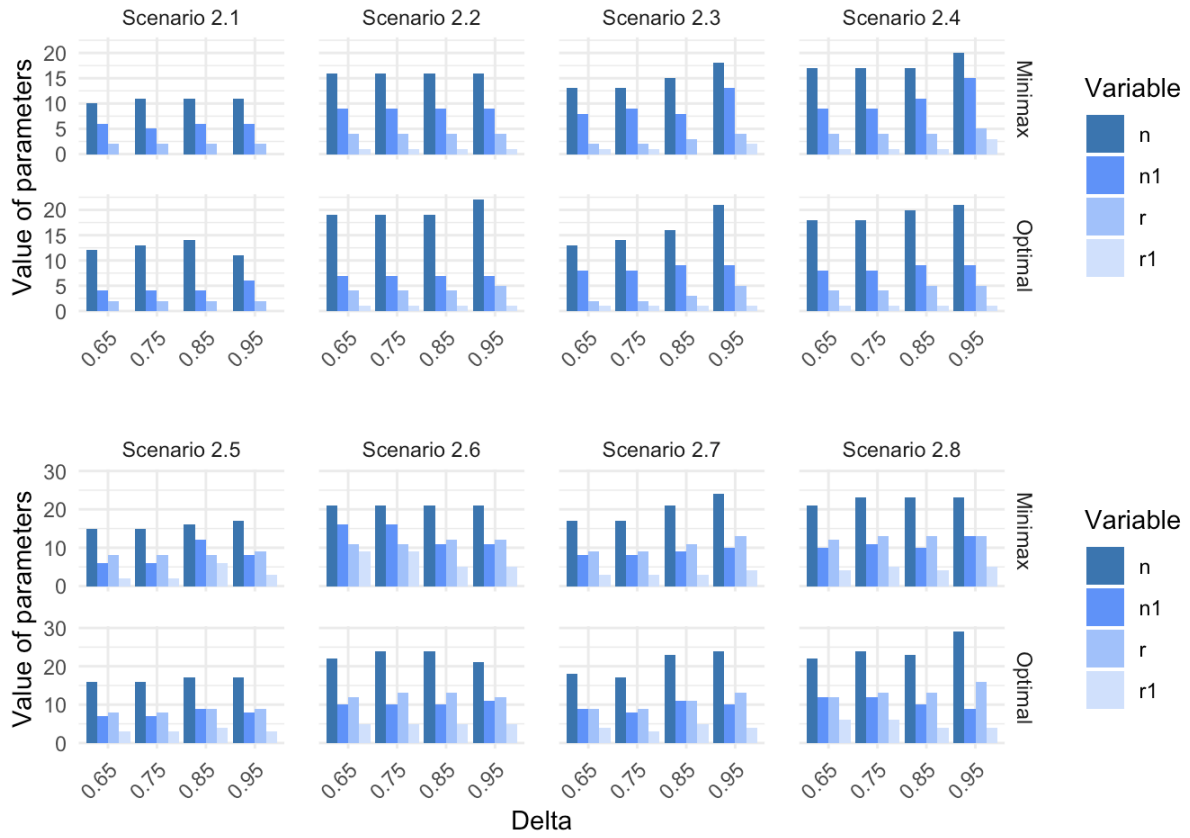

FIGURE 1 Design parameters for Optimal-IBPW and Minimax-IBPW by delta

4 | DISCUSSION

Single-arm trials are prevalent in phase II oncology trials primarily due to their efficiency, despite well-known concerns such as patient selection affecting the accuracy of efficacy assessments. Randomized trials are recognized as the gold standard for establishing the effectiveness of new treatments because they ensure a fair comparison between treatment arms. However, such trials often require large sample sizes that are not feasible in the early phase settings. Current randomized phase II trial designs involve some compromises; for instance, the highly efficient SWE[10] pick-the-winner design lacks control of the type I error, while Simon's two-stage design based multi-arm trials control the type I and II errors for treatment versus historical control comparisons but typically do not attempt between-arm comparisons due to lack of power. Consequently, the debate regarding the use of single-arm versus randomized designs for phase II oncology trials has been ongoing[17,18]. There is a need to further improve the designs of randomized phase II oncology trials.

The BPW design[14] represents an attractive development, extending Simon's two-stage based multi-arm design by incorporating a winner selection strategy. This design controls the type I and II errors for comparisons between treatment and the historical control and allows a limited two-arm comparison through winner selection. It also calculates the "overall power" and "overall type I error" for winner selection. Although the design is highly efficient, it is typically underpowered for winner selection and frames the hypotheses within a limited single-arm or two-arm response rate comparison framework.

In this paper, we propose the IBPW design, inspired by the BPW design. Our method includes a new hypotheses specification framework aimed at winner selection rather than direct response rate comparison. This hypotheses specification framework can be seen as an extension of the hypotheses specification in the SWE[10] design by including the null portion. It is more flexible and enables the testing of a broader range of clinically meaningful hypotheses, as demonstrated through examples described in our methods and simulation sections. Coupled with a randomized two-stage design and a winner selection strategy like that in the BPW approach, we have developed algorithms, along with corresponding R code and a Shiny App, to identify efficient designs with optimal parameters that tightly controls the type I and II errors for winner selection. Our simulations show that our IBPW designs are similarly efficient as the BPW, or other Simon's two-stage based randomized designs. They feature optimal sample sizes and decision cut-offs and tightly control prespecified type I and II error rates for winner selection, which cannot be readily achieved with existing approaches.

Our IBPW approach offers a pathway for further development of efficient winner-selection-based designs that tightly control both type I and II errors. Firstly, the endpoint that our current IBPW design focuses on is binary. There are situations where other types of endpoints are preferred. For example, a continuous endpoint, such as the values of PSA or other tumor biomarkers, or the numbers of circulating tumor cells (CTCs), may be good measures of the anti-tumor activity of some study agents. A time-to-event endpoint, such as progression-free survival (PFS), can be appropriate for treatments of advanced diseases, or ordinal or multivariate endpoints may be preferred in trials involving novel molecular targeted agents. Secondly, our current approach employs a winner selection strategy that includes a two-stage screening and a limited Bayesian two-group comparison. It is possible to extend our approach by considering different winner selection strategies with various number of study stages.

ACKNOWLEDGEMENT

The research of XKZ was partially supported by the National Center for Advancing Translational Sciences of the National Institutes of Health Grant UL1TR002384.

References

1. Wong CH, Siah KW, Lo AW. Estimation of clinical trial success rates and related parameters. *Biostatistics.* 2019;20:273-286.

2. Pan H, Yuan Y. *Bayesian Adaptive Design for Immunotherapy and Targeted Therapy.* Singapore:Springer; 2023. 91–118 p.

3. George SL. Response rate as an endpoint in clinical trials. *JNCI.* 2007;99:98-99.

4. Ritchie G, Gasper H, Man J, Lord S, Marschner I, Friedlander M, Lee CK. Defining the most appropriate primary end point in phase II trials of immune checkpoint inhibitors for advanced solid cancers: a systematic review and meta-analysis. *JAMA oncology.* 2018;4:522-528.

5. Zhao Y, Li D, Liu R, Yuan Y. Bayesian optimal phase II designs with dual-criterion decision making. *Pharm Stat.* 2023;22:605-618.

6. Gehan EA. The determination of the number of patients required in a follow-up trial of a new chemotherapeutic agent. *J Chronic Dis.* 1961;13:346–353.

7. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials.* 1989;10:1–10.

8. Ivanova A, Paul B, Marchenko O, Song G, Patel N, Moschos SJ. Nine-year change in statistical design, profile, and success rates of phase II oncology trials. *J Biopharm Stat.* 2016;26:141-149.

9. Chen Y, Chen Z, Mori M. A new statistical decision rule for single-arm phase II oncology trials. *Stat Methods Med Res.* 2016;25:118-132.

10. Simon R, Wittes RE, Ellenberg SS. Randomized phase II clinical trials. *Cancer Treat Rep.* 1985;69:1375-1381.

11. Lee JJ, Feng L. Randomized phase II designs in cancer clinical trials: current status and future directions. *J Clin Oncol.* 2005;23:4450-4457.

12. Steinberg SM, Venzon DJ. Early selection in a randomized phase II clinical trial. *Stat Med.* 2002;21:1711-1726.

13. Zhou H, Chen C, Sun L, Yuan Y. Bayesian optimal phase II clinical trial design with time-to-event endpoint. *Pharm. Stat.* 2020;19:776-786.

14. Chen DT, Huang PY, Lin HY, Chiappori AA, Gabrilovich DI, Haura EB, ... Gray JE. A Bayesian pick-the-winner design in a randomized phase II clinical trial. *Oncotarget.* 2017;8:88376–88385.

15. Pham-Gia T, Turkkan N, Eng P. Bayesian analysis of the difference of two proportions. *Commun. Stat. - Theory Methods.* 1993;22:1755-1771.

16. Yang R, Berger JO. *A catalog of noninformative priors*. 1998; http://www.stats.org.uk/priors/noninformative/YangBerger1998.pdf.

17. Buyse M. Randomized designs for early trials of new cancer treatments—an overview. *Drug Inf. J.: DIJ/Drug Information Association.* 2000;34: 387-396.

18. Grayling MJ, Dimairo M, Mander AP, Jaki TF. A review of perspectives on the use of randomization in phase II oncology trials. *JNCI.* 2019;111: 1255-1262.

TABLE 1 Comparison between optimal/minimax design obtained by BPW and IBPW

	n	n_1	r	r_1	$EN(p_0)$	$power$	$alpha$
Scenario 1.1	$H_0: p_{B_0}=0.05$ v.s $p_{A_0}=0.05$, $H_1: p_{B_1}=0.4$ v.s $p_{A_1}=0.05$						
optimal-BPW	8	4	1	0	9.484	0.805	0.044
optimal-IBPW	8	4	1	0	9.484	0.805	0.044
minimax-BPW	7	5	1	0	10.905	0.805	0.041
minimax-IBPW	7	5	1	0	10.905	0.805	0.041
Scenario 1.2	$H_0: p_{B_0}=0.05$ v.s $p_{A_0}=0.05$, $H_1: p_{B_1}=0.35$ v.s $p_{A_1}=0.05$						
optimal-BPW	11	4	1	0	10.597	0.780	0.061
optimal-IBPW	9	5	1	0	11.810	0.800	0.057
minimax-BPW	8	6	1	0	13.060	0.791	0.052
minimax-IBPW	9	5	1	0	11.810	0.800	0.057
Scenario 1.3	$H_0: p_{B_0}=0.1$ v.s $p_{A_0}=0.1$, $H_1: p_{B_1}=0.4$ v.s $p_{A_1}=0.1$						
optimal-BPW	11	4	2	0	12.815	0.762	0.067
optimal-IBPW	14	4	2	0	14.878	0.804	0.100
minimax-BPW	10	5	2	0	14.095	0.777	0.061
minimax-IBPW	11	6	2	0	16.686	0.811	0.078
Scenario 1.4	$H_0: p_{B_0}=0.2$ v.s $p_{A_0}=0.2$, $H_1: p_{B_1}=0.5$ v.s $p_{A_1}=0.2$						
optimal-BPW	13	6	4	1	16.825	0.766	0.078
optimal-IBPW	17	6	5	1	19.582	0.815	0.076
minimax-BPW	12	8	4	1	19.973	0.758	0.067
minimax-IBPW	13	8	4	1	20.967	0.800	0.088
Scenario 1.5	$H_0: p_{B_0}=0.2$ v.s $p_{A_0}=0.2$, $H_1: p_{B_1}=0.4$ v.s $p_{A_1}=0.2$						
optimal-BPW	25	12	7	2	35.483	0.762	0.090
optimal-IBPW	30	15	8	3	40.555	0.800	0.097
minimax-BPW	24	14	7	2	39.039	0.756	0.080
minimax-IBPW	29	16	8	3	42.449	0.800	0.090
Scenario 1.6	$H_0: p_{B_0}=0.4$ v.s $p_{A_0}=0.4$, $H_1: p_{B_1}=0.6$ v.s $p_{A_1}=0.4$						
optimal-BPW	38	12	18	5	41.409	0.763	0.089
optimal-IBPW	38	15	18	6	47.949	0.809	0.100
minimax-BPW	28	16	14	6	43.348	0.742	0.089
minimax-IBPW	35	14	17	5	49.594	0.802	0.094
Scenario 1.7	$H_0: p_{B_0}=0.5$ v.s $p_{A_0}=0.5$, $H_1: p_{B_1}=0.7$ v.s $p_{A_1}=0.5$						
optimal-BPW	32	12	19	6	39.488	0.754	0.082
optimal-IBPW	38	17	22	9	47.210	0.809	0.091
minimax-BPW	28	15	17	7	43.000	0.753	0.082
minimax-IBPW	34	22	20	12	50.282	0.800	0.090
Scenario 1.8	$H_0: p_{B_0}=0.6$ v.s $p_{A_0}=0.6$, $H_1: p_{B_1}=0.8$ v.s $p_{A_1}=0.6$						
optimal-BPW	31	11	21	7	33.851	0.766	0.090
optimal-IBPW	29	12	20	7	38.898	0.802	0.095
minimax-BPW	24	11	17	6	35.852	0.744	0.085
minimax-IBPW	29	12	20	7	38.898	0.802	0.095

TABLE 2 Results of the IBPW design under eight different scenarios

	n	n_1	r	r_1	$EN(p_0)$	$power$	$alpha$
Scenario 2.1	$H_0: p_{B_0} = 0.1$ v.s $p_{A_0} = 0.1$, $H_1: p_{B_1} = 0.4$ v.s $p_{A_1} = 0.1$						
Optimal	14	4	2	0	14.878	0.804	0.100
Minimax	11	6	2	0	16.686	0.811	0.078
Scenario 2.2	$H_0: p_{B_0} = 0.15$ v.s $p_{A_0} = 0.1$, $H_1: p_{B_1} = 0.4$ v.s $p_{A_1} = 0.1$						
Optimal	19	7	4	1	19.197	0.804	0.098
Minimax	16	9	4	1	22.380	0.806	0.073
Scenario 2.3	$H_0: p_{B_0} = 0.1$ v.s $p_{A_0} = 0.1$, $H_1: p_{B_1} = 0.4$ v.s $p_{A_1} = 0.15$						
Optimal	18	8	3	1	19.738	0.803	0.062
Minimax	15	8	3	0	23.973	0.802	0.051
Scenario 2.4	$H_0: p_{B_0} = 0.15$ v.s $p_{A_0} = 0.1$, $H_1: p_{B_1} = 0.4$ v.s $p_{A_1} = 0.15$						
Optimal	20	9	5	1	24.882	0.813	0.060
Minimax	17	12	4	2	25.875	0.802	0.088
Scenario 2.5	$H_0: p_{B_0} = 0.4$ v.s $p_{A_0} = 0.4$, $H_1: p_{B_1} = 0.7$ v.s $p_{A_1} = 0.4$						
Optimal	17	9	9	4	22.265	0.806	0.071
Minimax	16	12	8	6	25.266	0.809	0.094
Scenario 2.6	$H_0: p_{B_0} = 0.45$ v.s $p_{A_0} = 0.4$, $H_1: p_{B_1} = 0.7$ v.s $p_{A_1} = 0.4$						
Optimal	24	10	13	5	25.989	0.806	0.091
Minimax	21	16	11	9	32.912	0.800	0.099
Scenario 2.7	$H_0: p_{B_0} = 0.4$ v.s $p_{A_0} = 0.4$, $H_1: p_{B_1} = 0.7$ v.s $p_{A_1} = 0.45$						
Optimal	22	9	11	4	24.931	0.810	0.080
Minimax	19	10	10	4	26.604	0.803	0.075
Scenario 2.8	$H_0: p_{B_0} = 0.45$ v.s $p_{A_0} = 0.4$, $H_1: p_{B_1} = 0.7$ v.s $p_{A_1} = 0.45$						
Optimal	24	12	13	6	29.027	0.808	0.097
Minimax	23	11	13	5	29.361	0.804	0.081

TABLE 3 Operating characteristics of Optimal and Minimax IBPW designs

Optimal design

Scenario	p_{B_0}	p_{A_0}	p_{B_1}	p_{A_1}	$2n_1\|H_0$	$n_1+n\|H_0$	$2n\|H_0$	$2n_1\|H_1$	$n_1+n\|H_1$	$2n\|H_1$	$Y_A>r\|p_{A_0}$	$Y_B>r\|p_{B_0}$	$Y_A>r\|p_{A_1}$	$Y_B>r\|p_{B_1}$
2.1	0.1	0.1	0.4	0.1	0.430	0.452	0.119	0.085	0.616	0.299	0.112	0.112	0.112	0.852
2.2	0.15	0.1	0.4	0.1	0.609	0.348	0.042	0.135	0.740	0.127	0.024	0.100	0.024	0.813
2.3	0.1	0.1	0.4	0.15	0.661	0.304	0.035	0.070	0.624	0.307	0.066	0.066	0.197	0.882
2.4	0.15	0.1	0.4	0.15	0.465	0.445	0.090	0.042	0.585	0.372	0.010	0.061	0.061	0.844
2.5	0.4	0.4	0.7	0.4	0.538	0.392	0.072	0.072	0.688	0.241	0.077	0.077	0.077	0.849
2.6	0.45	0.4	0.7	0.4	0.616	0.341	0.043	0.125	0.732	0.141	0.036	0.094	0.036	0.821
2.7	0.4	0.4	0.7	0.45	0.538	0.392	0.072	0.061	0.597	0.341	0.087	0.087	0.182	0.884
2.8	0.45	0.4	0.7	0.45	0.622	0.336	0.041	0.087	0.683	0.230	0.039	0.101	0.101	0.852

Minimax design

Scenario	p_{B_0}	p_{A_0}	p_{B_1}	p_{A_1}	$2n_1\|H_0$	$n_1+n\|H_0$	$2n\|H_0$	$2n_1\|H_1$	$n_1+n\|H_1$	$2n\|H_1$	$Y_A>r\|p_{A_0}$	$Y_B>r\|p_{B_0}$	$Y_A>r\|p_{A_1}$	$Y_B>r\|p_{B_1}$
2.1	0.1	0.1	0.4	0.1	0.282	0.498	0.220	0.025	0.528	0.446	0.085	0.085	0.085	0.866
2.2	0.15	0.1	0.4	0.1	0.465	0.446	0.090	0.055	0.736	0.210	0.016	0.074	0.016	0.815
2.3	0.1	0.1	0.4	0.15	0.185	0.490	0.324	0.005	0.280	0.715	0.054	0.054	0.174	0.905
2.4	0.15	0.1	0.4	0.15	0.654	0.317	0.029	0.061	0.697	0.242	0.020	0.090	0.090	0.852
2.5	0.4	0.4	0.7	0.4	0.709	0.266	0.026	0.099	0.761	0.140	0.105	0.105	0.105	0.867
2.6	0.45	0.4	0.7	0.4	0.825	0.169	0.007	0.165	0.787	0.048	0.044	0.103	0.044	0.819
2.7	0.4	0.4	0.7	0.45	0.401	0.466	0.135	0.024	0.504	0.472	0.081	0.081	0.171	0.897
2.8	0.45	0.4	0.7	0.45	0.477	0.434	0.091	0.050	0.612	0.338	0.031	0.084	0.084	0.847

TABLE 4 Impact of changes in hypotheses on the operating characteristics of optimal and minimax IBPW designs

Designs	p_{A_0}	p_{B_0}	p_{A_1}	p_{B_1}	$power$	$alpha$	$EN(p_0)$
Scenario 2.4-Optimal	0.1	0.15	0.15	0.4	0.813	0.06	24.882
($n = 20$, $n_1 = 9$, $r = 5$, $r_1 = 1$)	0.1	0.1	0.15	0.4	0.813	0.01	22.953
	0.1	0.15	0.1	0.4	0.839	0.06	24.882
	0.1	0.1	0.1	0.4	0.839	0.01	22.953
Scenario 2.4-Minimax	0.1	0.15	0.15	0.4	0.802	0.088	25.875
($n = 17$, $n_1 = 12$, $r = 4$,$r_1 = 2$)	0.1	0.1	0.15	0.4	0.802	0.02	25.109
	0.1	0.15	0.1	0.4	0.842	0.088	25.875
	0.1	0.1	0.1	0.4	0.842	0.02	25.109
Scenario 2.8-Optimal	0.4	0.45	0.45	0.7	0.808	0.097	29.027
($n = 24$, $n_1 = 12$, $r = 13$, $r_1 = 6$)	0.4	0.4	0.45	0.7	0.808	0.038	27.797
	0.4	0.45	0.4	0.7	0.836	0.097	29.027
	0.4	0.4	0.4	0.7	0.836	0.038	27.797
Scenario 2.8-Minimax	0.4	0.45	0.45	0.7	0.804	0.081	29.361
($n = 23$, $n_1 = 11$, $r = 13$, $r_1 = 5$)	0.4	0.4	0.45	0.7	0.804	0.03	27.916
	0.4	0.45	0.4	0.7	0.832	0.081	29.361
	0.4	0.4	0.4	0.7	0.832	0.03	27.916

APPENDIX

SUPPLEMENTARY TABLES

Table S1: Probabilities of all possible winner selection outcomes for Optimal IBPW design as listed in Table 2.

H_0: $p_{B_0} = 0.1$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.4$ v.s $p_{A_1} = 0.1$			
Scenario 2.1 $AB.ES.H0^{(1)}$: 0.882		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.430	0.152	0.074	A.fail.stage1	0.085	0.012	0.559
	A.fail.stage2	0.152	0.054	0.026	A.fail.stage2	0.030	0.004	0.197
	A.pass	0.074	0.026	0.013($0^{(2)}$)	A.pass	0.015	0.002	0.096(0.047)
H_0: $p_{B_0} = 0.15$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.4$ v.s $p_{A_1} = 0.1$			
Scenario 2.2 $AB.ES.H0$: 0.958		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.609	0.156	0.085	A.fail.stage1	0.135	0.024	0.692
	A.fail.stage2	0.090	0.023	0.013	A.fail.stage2	0.020	0.004	0.103
	A.pass	0.017	0.004	0.002(0)	A.pass	0.004	0.001	0.019(0.01)
H_0: $p_{B_0} = 0.1$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.4$ v.s $p_{A_1} = 0.15$			
Scenario 2.3 $AB.ES.H0$: 0.965		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.661	0.098	0.054	A.fail.stage1	0.070	0.007	0.580
	A.fail.stage2	0.098	0.015	0.008	A.fail.stage2	0.016	0.002	0.129
	A.pass	0.054	0.008	0.004(0)	A.pass	0.021	0.002	0.174(0.094)
H_0: $p_{B_0} = 0.15$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.40$ v.s $p_{A_1} = 0.15$			
Scenario 2.4 $AB.ES.H0$: 0.910		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.465	0.263	0.047	A.fail.stage1	0.042	0.051	0.506
	A.fail.stage2	0.129	0.073	0.013	A.fail.stage2	0.024	0.029	0.287
	A.pass	0.006	0.003	0.001(0)	A.pass	0.004	0.005	0.051(0.02)
H_0: $p_{B_0} = 0.4$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.4$			
Scenario 2.5 $AB.ES.H0$: 0.929		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.538	0.139	0.057	A.fail.stage1	0.072	0.038	0.623
	A.fail.stage2	0.139	0.036	0.015	A.fail.stage2	0.019	0.010	0.161
	A.pass	0.057	0.015	0.006(0)	A.pass	0.008	0.004	0.006(0.022)
H_0: $p_{B_0} = 0.45$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.4$			
Scenario 2.6 $AB.ES.H0$: 0.957		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.616	0.140	0.078	A.fail.stage1	0.125	0.024	0.684
	A.fail.stage2	0.096	0.022	0.012	A.fail.stage2	0.019	0.004	0.106
	A.pass	0.027	0.006	0.003(0)	A.pass	0.005	0.001	0.030(0.016)
H_0: $p_{B_0} = 0.4$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.45$			
Scenario 2.7 $AB.ES.H0$: 0.929		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.538	0.132	0.064	A.fail.stage1	0.061	0.010	0.55
	A.fail.stage2	0.132	0.032	0.016	A.fail.stage2	0.019	0.003	0.174
	A.pass	0.064	0.016	0.008(0)	A.pass	0.018	0.003	0.161(0.087)
H_0: $p_{B_0} = 0.45$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.45$			
Scenario 2.8 $AB.ES.H0$: 0.959		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
	A.fail.stage1	0.622	0.134	0.085	A.fail.stage1	0.087	0.022	0.630
	A.fail.stage2	0.088	0.019	0.012	A.fail.stage2	0.019	0.005	0.136
	A.pass	0.029	0.006	0.004(0)	A.pass	0.012	0.003	0.086(0.042)

(1) $AB.ES.H0$ represents the probability of arm A or B early stopping under H_0

(2) The number in parentheses represents the posterior probability of arm B being claimed as the winner

Table S2: Probabilities of all possible winner selection outcomes for Minimax IBPW design as listed in Table 2.

H_0: $p_{B_0} = 0.1$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.4$ v.s $p_{A_1} = 0.1$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.1	A.fail.stage1	0.282	0.204	0.045	A.fail.stage1	0.025	0.046	0.46
$AB.ES.H0^{(1)}$: 0.780	A.fail.stage2	0.204	0.147	0.033	A.fail.stage2	0.018	0.033	0.332
	A.pass	0.045	0.033	0.007($0^{(2)}$)	A.pass	0.004	0.007	0.074(0.018)
H_0: $p_{B_0} = 0.15$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.4$ v.s $p_{A_1} = 0.1$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.2	A.fail.stage1	0.465	0.253	0.058	A.fail.stage1	0.055	0.089	0.631
$AB.ES.H0$: 0.910	A.fail.stage2	0.125	0.068	0.016	A.fail.stage2	0.015	0.024	0.171
	A.pass	0.010	0.005	0.001(0)	A.pass	0.001	0.002	0.013(0.004)
H_0: $p_{B_0} = 0.1$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.4$ v.s $p_{A_1} = 0.15$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.3	A.fail.stage1	0.185	0.222	0.023	A.fail.stage1	0.005	0.021	0.247
$AB.ES.H0$: 0.676	A.fail.stage2	0.222	0.265	0.028	A.fail.stage2	0.009	0.043	0.501
	A.pass	0.023	0.028	0.003(0)	A.pass	0.003	0.014	0.157(0.055)
H_0: $p_{B_0} = 0.15$ v.s $p_{A_0} = 0.1$					H_1: $p_{B_1} = 0.40$ v.s $p_{A_1} = 0.15$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.4	A.fail.stage1	0.654	0.155	0.080	A.fail.stage1	0.061	0.047	0.627
$AB.ES.H0$: 0.971	A.fail.stage2	0.067	0.016	0.008	A.fail.stage2	0.015	0.011	0.148
	A.pass	0.015	0.003	0.002(0)	A.pass	0.008	0.006	0.077(0.026)
H_0: $p_{B_0} = 0.4$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.4$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.5	A.fail.stage1	0.709	0.045	0.088	A.fail.stage1	0.099	0.013	0.730
$AB.ES.H0$: 0.975	A.fail.stage2	0.045	0.003	0.006	A.fail.stage2	0.006	0.001	0.046
	A.pass	0.088	0.006	0.011(0)	A.pass	0.012	0.002	0.091(0.033)
H_0: $p_{B_0} = 0.45$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.4$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.6	A.fail.stage1	0.825	0.020	0.097	A.fail.stage1	0.165	0.005	0.771
$AB.ES.H0$: 0.993	A.fail.stage2	0.013	0.000	0.001	A.fail.stage2	0.003	0.000	0.012
	A.pass	0.039	0.001	0.005(0)	A.pass	0.008	0.000	0.036(0.017)
H_0: $p_{B_0} = 0.4$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.45$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.7	A.fail.stage1	0.401	0.181	0.052	A.fail.stage1	0.024	0.028	0.453
$AB.ES.H0$: 0.865	A.fail.stage2	0.181	0.082	0.023	A.fail.stage2	0.015	0.018	0.292
	A.pass	0.052	0.023	0.007(0)	A.pass	0.008	0.009	0.153(0.058)
H_0: $p_{B_0} = 0.45$ v.s $p_{A_0} = 0.4$					H_1: $p_{B_1} = 0.7$ v.s $p_{A_1} = 0.45$			
		B.fail.stage1	B.fail.stage2	B.pass		B.fail.stage1	B.fail.stage2	B.pass
Scenario 2.8	A.fail.stage1	0.477	0.214	0.063	A.fail.stage1	0.050	0.047	0.536
$AB.ES.H0$: 0.910	A.fail.stage2	0.137	0.061	0.018	A.fail.stage2	0.022	0.021	0.240
	A.pass	0.020	0.009	0.003(0)	A.pass	0.007	0.006	0.071(0.028)

(1) $AB.ES.H0$ represents the probability of arm A or B early stopping under H_0

(2) The number in parentheses represents the posterior probability of arm B being claimed as the winner