The New England Journal of Statistics in Data Science Volume 3, 272–281 (2025)
DOI: https://doi.org/10.51387/25-NEJSDS83

Three-Outcome Dual-Criterion Randomized Phase II Clinical Trial Design

Yujia WANG*, Xiaohan CHI*, and Ruitao LIN

Abstract

The high cost of drug development and the relatively low success rates of phase III clinical trials highlight the need for improved and reasonably sized phase II trial designs, especially when responses observed in treatment and control could not lead to a clear-cut decision warranting further studies. To this end, we propose a three-outcome dual-criterion randomized (TDR) trial design, which implements inconclusive region sculpting using boundaries defined by both statistically significant differences between treatment and control as well as the clinical relevance of treatment responses. We provide statistical justifications for the TDR design in both one-stage and two-stage trial settings. Additionally, we evaluate its operating characteristics through a comparison with existing designs. The proposed design is shown able to achieve sample size saving and type II error reduction while controlling the type I error at a marginal cost of power reduction. Lastly, robustness under various deviations from the assumed control response rate is also demonstrated.

keywords and phrases: Phase II trial design, Inconclusive region, Clinical relevance, Binary outcome, Sample size.

1. INTRODUCTION

With the recent development in cancer treatment and regulations, a greater focus has been placed on randomized designs in phase II cancer clinical trials. In general, phase II trials are a vital step in oncology drug development. A phase II clinical trial should screen out inefficacious agents while warranting subsequent large-scale phase III clinical trials when a treatment demonstrates safety and efficacy [9, 13]. However, the rate of success in phase III trials is generally low [7], highlighting the need for more careful decision-making in phase II trials [17]. A plethora of trial designs have been proposed for phase II trials, including the commonly used Simon's two-stage design [19], historical controls [1], the reference control arm design [5], the pick-the-winner design [18], and the screening design [14]. However, these methods, especially those that utilize a single-armed design, are subject to potential biases due to shifts in patient selection and evolving standards of care [13]. Conventional hypothesis testing results in two possible outcomes: either rejecting or accepting the null hypothesis, which often poses a dilemma for investigators, especially when the observed responses fall near the decision boundaries. In such scenarios, the dichotomous framework requires a definitive acceptance or rejection of one hypothesis over the other, despite the inherent uncertainty in observed data. Considering the high cost of drug development and long development phases [21], the impact of incorrect decisions can be substantial. In view of the issue, several three-outcome designs have been proposed. In a phase II trial setting, Storer [20] proposed a single-armed three-outcome design that allows for rejecting neither $H_0 : p \leq p_1$ nor $H_a : p \geq p_2$ when observed response rates fall between probabilities p_1 and p_2. This design optimizes the sample size to meet constraints on the probability of rejecting neither hypothesis. Sargent et al. [15] proposed an alternative three-outcome design with an inconclusive region defined by two cutoff points for observed responses. Building on this, Hong and Wang [6] extended Sargent's design to a two-armed randomized controlled trial that controls design error rates and inconclusiveness probabilities, resulting in considerable sample size savings compared to traditional two-outcome designs.

Concurrently, researchers seek to enhance the validity and practicality of phase II trials by incorporating a second criterion of clinical relevance in decision rules. Fisch et al. [4] raised the question of whether statistical significance between treatment and control arms alone is sufficient to justify advancing to phase III trials, noting that a statistically significant but minor improvement may not warrant further investment. Thus, they proposed a proof-of-concept design where dual criteria of significance and relevance were evaluated. Subsequently, Litwin et al. [10] extended this approach to a two-stage randomized controlled trial method. They proposed early termination probabilities under the null and alternative hypotheses to derive the stage 1 sample size and employed an incremental search for stage 2 sample size determination. This method showed substantial sample size savings, yet we see merits in addressing the borderline response rates to further reduce false positives.

In this paper, we propose a randomized controlled phase II clinical trial design that considers both the uncertainty in

*Corresponding authors.

trial outcomes and clinical relevance. By incorporating the inconclusive regions in the hypothesis testing framework, the proposed design allows practical considerations such as clinical, regulatory, and commercial decision-making. The adoption of the dual criteria further ensures the predicted power to warrant a phase III trial and reduces the type I error. The remainder of the paper is organized as follows. In Section 2, we propose three-outcome dual-criterion randomized designs with controls on the inconclusive region, presented in both one-stage and two-stage manners. We also describe the sample size determination procedure and introduce a loss function for optimizing the design parameters. In Section 3, we evaluate the proposed method numerically and compare it with existing methods. In Section 4, we apply the proposed design to data from the VIT-0910 trial. The paper is concluded with discussions in Section 5. The TDR sample size calculation program in the form of `R` code is available online at https://github.com/ywangaz/TDRdesign.

2. METHODS

In this section, we describe a three-outcome dual-criterion randomized (TDR) design for phase II trials with binary efficacy endpoints. The design aims to attain sample size savings while controlling type I and type II errors as well as maintaining adequate statistical power. This is achieved by sculpting the hypothesis rejection region, taking both statistical significance and clinical relevance into account. The probability of incorrectly rejecting either the null hypothesis or the alternative hypothesis is reduced by introducing an inconclusive region, which allows comprehensive considerations of other aspects in addition to statistical significance and clinical relevance in drug development when observed results are borderline. We first focus on the TDR design in a one-stage trial setting, and the design in a two-stage trial setting will also be discussed. For simplicity, the 1:1 randomization is illustrated in this paper, and the design can be applied to other randomization schemes where appropriate.

2.1 TDR One-Stage Design

In phase II trials with binary efficacy endpoints, let p_E and p_C denote the true response rates for the experimental arm and the control arm, and $\hat{p}_E$ and $\hat{p}_C$ denote the observed response rates. We aim to address the two-sample test with the following hypotheses:

$$H_0 : p_E = p_C = p_0,$$
$$H_a : p_E = p_1,\ p_C = p_0,\ p_1 > p_0.$$

Traditional approaches for the two-sample test rely solely on between-arm comparisons and yield only binary trial outcomes: either rejecting or accepting H_0. To reduce the required sample size and draw more meaningful conclusions from two-sample tests, we combine one-sample rejection decision rules with traditional two-sample rules and introduce a statistically inconclusive region.

In a one-stage setting, assume n_E patients are recruited to the experimental arm and n_C patients are recruited to the control arm, leading to a total of $N = n_E + n_C$ patients. Let y_E and y_C denote the number of patients demonstrating successful responses in the experimental and control arm, respectively. The decision rules for reaching one of the three outcomes of rejecting H_0, rejecting H_a, or rejecting neither are defined as follows:

$$\begin{aligned}
&\text{If } \hat{p}_E - \hat{p}_C \geq p_s \cap \hat{p}_E \geq p_m, \text{ reject } H_0;\\
&\text{If } \hat{p}_E - \hat{p}_C < p_s, \text{ reject } H_a;\\
&\text{If } \hat{p}_E - \hat{p}_C \geq p_s \cap \hat{p}_E < p_m, \text{ declare statistically inconclusive,}
\end{aligned} \tag{2.1}$$

where p_s denotes the statistical significance boundary ($p_s > 0$), and p_m denotes the clinical relevance boundary ($p_m > 0$). Given a 1:1 randomization, the decision rules in Equation (2.1) can be simplified by assuming $n_E = n_C = N/2$, that is

$$\begin{aligned}
&\text{If } y_E - y_C \geq s \cap y_E \geq m, \text{ reject } H_0;\\
&\text{If } y_E - y_C < s, \text{ reject } H_a;\\
&\text{If } y_E - y_C \geq s \cap y_E < m, \text{ declare statistically inconclusive,}
\end{aligned}$$

where $s = \frac{N}{2}p_s$, and $m = \frac{N}{2}p_m$. Here we refer to this design as a 2-by-2 TDR design, where the decision rules for both statistical significance and clinical relevance each contain two regions: $y_E - y_C \geq s$ or $y_E - y_C < s$ for statistical significance, and $y_E \geq m$ or $y_E < m$ for clinical relevance. Under the independent and normality assumption, the conditions of the decision rules are equivalent to constructing two z-test statistics.

The statistically inconclusive region is reserved for the situation where there are substantial differences between the experimental arm and the control, while the observed responses in the experimental arm are suboptimal in terms of the historical control rate. This may occur when the trial population differs from the population used to derive the historical control rate. In this case, the clinical decision regarding whether to proceed to a phase III trial or terminate the current trial requires more deliberation. Primary investigators and statisticians should comprehensively review factors such as regulatory requirements, commercial potential, and practicality of administering the treatment, in order to decide on warranting further investigations.

In our one-stage TDR design, we use exact binomial probabilities to calculate the type I error α, type II error β, and inconclusive region probabilities η under H_0 and γ under H_a, as follows:

$$\alpha = \sum_{D_1 \cap D_2} B(y_E, n_E, p_E \mid H_0) B(y_C, n_C, p_C \mid H_0),$$
$$\beta = \sum_{D_1'} B(y_E, n_E, p_E \mid H_a) B(y_C, n_C, p_C \mid H_a),$$

$$\eta = \sum_{D_1 \cap D_2'} B(y_E, n_E, p_E \mid H_0) B(y_C, n_C, p_C \mid H_0),$$
$$\gamma = \sum_{D_1 \cap D_2'} B(y_E, n_E, p_E \mid H_a) B(y_C, n_C, p_C \mid H_a),$$

where $D_1 = \{(y_E, y_C) : y_E - y_C \geq s\}$, $D_2 = \{y_E : y_E \geq m\}$, and D_1' and D_2' denote the complementary sets of D_1 and D_2. In addition, we introduce $\lambda = (\eta + \gamma)/2$ as the expected inconclusive probability. This definition is based on a common assumption that the unknown true response rates of both arms vary uniformly between the null and alternative hypotheses. The statistically inconclusive regions under H_0 and H_a can be controlled simultaneously by constraining γ and λ instead of γ and η to prevent highly imbalanced inconclusive regions under H_0 and H_a. To provide finer control over inconclusive regions, it is possible to introduce both upper and lower boundaries for determining the statistical significance. This extension is referred to as a 3-by-2 TDR design, with three regions for statistical significance (i.e. $y_E - y_C \geq s$, $r < y_E - y_C < s$, or $y_E - y_C \leq r$) and two regions for clinical relevance (i.e. $y_E \geq m$ or $y_E < m$) in the decision rules. A detailed description is provided in Section 1.1 of the Supplementary Materials.

2.2 TDR Two-Stage Design

In this section, we consider extending the proposed design to a two-stage trial setting, for the purpose of ethically stopping a trial early given insufficient evidence of efficacy. In stage 1, we enroll and randomize n_{C1} patients to the control arm and n_{E1} patients to the experimental arm. If the trial is not stopped early, we proceed to stage 2, where n_{C2} and n_{E2} patients are randomized to each arm. We denote $N_1 = n_{C1} + n_{E1}$ and $N_2 = n_{C2} + n_{E2}$ as the total sample size in stages 1 and 2, respectively. The number of responses observed in the stage 1 (or stage 2) are denoted as y_{E1} and y_{C1} (or y_{E2} and y_{C2}) for the treatment and control arms, respectively. At the conclusion of stage 1, an interim analysis is performed, and the trial proceeds to stage 2 if

$$y_{E1} - y_{C1} > s_1 \quad \text{and} \quad y_{E1} \geq m_1, \tag{2.2}$$

where s_1 is a statistical difference threshold for early stopping and m_1 is a clinical relevance threshold for early stopping. Note that the inconclusive regions are excluded in the interim analysis for simplicity. We denote the probabilities of proceeding to stage 2 under H_0 and H_a as $\Pr(S_1 \mid H_0)$ and $\Pr(S_1 \mid H_a)$, respectively, where $S_1 = \{n_{E1}, n_{C1}, s_1, m_1 : y_{E1} - y_{C1} > s_1 \cap y_{E1} \geq m_1\}$ represents the condition (2.2). In our design, we propose to control

$$\Pr(S_1 \mid H_0) \leq 1 - es_0 \quad \text{and} \quad \Pr(S_1 \mid H_a) \geq 1 - es_1,$$

where es_0 and es_1 are early stopping probability levels under the null and alternative hypotheses, respectively. In this paper, we set $es_0 = 0.50$ and $es_1 = 0.05$ as reasonable constraints.

In stage 2, the proposed design will proceed as described in Section 2.1, with type I error α, type II error β, and inconclusive region probabilities η under H_0 and γ under H_a defined as below:

$$\alpha = \sum_{D_1 \cap D_2} B(y_{E2}, n_{E2}, p_E \mid H_0) B(y_{C2}, n_{C2}, p_C \mid H_0) \Pr(S_1 \mid H_0),$$
$$\beta = \sum_{D_1'} B(y_{E2}, n_{E2}, p_E \mid H_a) B(y_{C2}, n_{C2}, p_C \mid H_a) \Pr(S_1 \mid H_a),$$
$$\eta = \sum_{D_1 \cap D_2'} B(y_{E2}, n_{E2}, p_E \mid H_0) B(y_{C2}, n_{C2}, p_C \mid H_0) \Pr(S_1 \mid H_0),$$
$$\gamma = \sum_{D_1 \cap D_2'} B(y_{E2}, n_{E2}, p_E \mid H_a) B(y_{C2}, n_{C2}, p_C \mid H_a) \Pr(S_1 \mid H_a),$$

where $D_1 = \{(y_{E2}, y_{C2}) : y_{E2} - y_{C2} \geq s_2 - (y_{E1} - y_{C1})\}$, $D_2 = \{y_{E2} : y_{E2} \geq m_2 - y_{E1}\}$, s_2 is the statistical significance boundary ($s_2 > y_{E1} - y_{C1}$), and m_2 is the clinical relevance boundary ($m_2 > y_{E1}$).

2.3 Sample Size Determination

With the introduction of the inconclusiveness region, the power of the TDR design is defined as $\pi = 1 - \beta - \gamma$, which is the probability of rejecting H_0 when H_a is true. Given a specific minimum target power $\pi_{\min}$ and maximum type II error level $\beta_{\max}$, we control the inconclusive region probabilities γ and λ under their maximum allowable constraints $\gamma_{\max}$ and $\lambda_{\max}$. As a result, there might be multiple sets of $(\alpha, \beta, \gamma, \lambda)$ satisfying these requirements. For example, given $\pi_{\min}$ and $\beta_{\max}$, the maximum allowable inconclusive probability under H_a is given by $\gamma_{\max} = 1 - (\beta_{\max} + \pi_{\min})$. To ensure proper control of the trial's inconclusive probabilities, a possible γ should not exceed this threshold (i.e., $\gamma \leq \gamma_{\max}$). Similarly, $\lambda_{\max}$ is given by $\lambda_{\max} = (\eta_{\max} + \gamma_{\max})/2$, where $\eta_{\max}$ represents the maximum target level for the inconclusive probability η. To facilitate parameter search under these constraints, we propose a loss function for systematic evaluation, which is discussed in Section 2.4.

The sample size for the proposed one-stage TDR design is determined using a two-step approach. Firstly, for each candidate sample size, sets of parameters $\{s, m, n_C, n_E\}$ are obtained through an incremental search over a grid of design parameters $\{(\alpha, \beta, \gamma, \lambda, \pi) : \alpha \leq \alpha_{\max}, \beta \leq \beta_{\max}, \gamma \leq \gamma_{\max}, \lambda \leq \lambda_{\max}, \pi \geq \pi_{\min}\}$, where $\alpha_{\max}$ denotes the maximum target type I error level. For each given set of $(\alpha, \beta, \gamma, \lambda, \pi)$ satisfying these constraints, we search for all possible pairs of s and m within pre-defined searching regions. Reasonable regions for s and m can be

$$\big\{s : s \in \big[p_C(n_E - kn_C), p_E n_E - k p_C n_C + 4\sigma_{\text{pool}}\big], \\ -n_C \leq s \leq n_E\big\}, \\ \big\{m : m \in [p_C n_E, p_E n_E + 4\sigma_E], m \leq n_E\big\},$$

where k is the randomization ratio, σ_{pool} is the pooled standard error, and σ_E is the standard error under H_a. The optimal (s, m) is then obtained by selecting the pair with

the smallest type I error α. Secondly, among all candidate sample sizes, the optimal design parameters $(\alpha, \beta, \gamma, \lambda, \pi)$ and the corresponding sample size are selected through a loss function that balances the trade-off between trial power and sample size. We provide a systematic evaluation of sample size and power using the loss function described in Section 2.4. The parameters yielding the smallest loss score are then selected as the optimal parameters.

The proposed one-stage design is expected to reduce type I error by introducing the inconclusive region, compared to the design by [10], at the same sample size. This occurs when the difference in the number of responses between the two arms is larger than s, but the number of observed responses in the experimental arm is less than m. Theoretically, given a fixed sample size N, the relationships between (s, m) and the associated error rates (or inconclusive probabilities) can be summarized as follows: when m is kept constant, increasing s reduces α, increases β, and reduces γ and η; when s is kept constant, increasing m reduces α, increases γ and η, and has no effect on β. Under H_0, the criterion of clinical relevance has less effect compared to its effect under H_a. This is because, when $p_E = p_C$, it is more likely that $y_E < m$. For the same reason, η is generally larger than γ.

2.4 Loss Function

Previous studies have established the practice of optimizing trial design using loss functions, such as [8, 12, 16], among others. In this design, we propose using a loss function to systematically evaluate the effects of inconclusive region constraints to optimize both sample size and power. For each sample size and its corresponding optimal design parameters $(\alpha, \beta, \gamma, \lambda, \pi)$, we calculate a loss score at the sample size n and power π with respect to a reference sample size n_0 and a reference power π_0 by a loss function $L(n, \pi, n_0, \pi_0)$. The reference sample size n_0 is the sample size per arm calculated using a standard two-group sample size calculation under the same hypothesis as the TDR design. The reference power π_0 is the corresponding power calculated in the reference sample size method, i.e. the probability of correctly accepting the alternative hypothesis of the reference sample size method. The primary principle of the loss function is to penalize an increase in sample size and a reduction in power. According to [11], a loss function should meet the following criteria:

1. Monotonicity: at a fixed sample size n or a fixed power π, the loss score should monotonously increase if power decreases or sample size increases.

$$L(n, \pi_1, n_0, \pi_0) > L(n, \pi_2, n_0, \pi_0) \Leftrightarrow \pi_1 < \pi_2;$$
$$L(n_1, \pi, n_0, \pi_0) > L(n_2, \pi, n_0, \pi_0) \Leftrightarrow n_1 > n_2.$$

2. Scale invariance: proportional scaling in sample size n and reference sample size n_0 at the same power, or proportional scaling in power π and reference power π_0 at the same sample size, should produce the same loss score.

$$L(n, \pi, n_0, \pi_0) = L(c \cdot n, \pi, c \cdot n_0, \pi_0);$$
$$L(n, \pi, n_0, \pi_0) = L(n, d \cdot \pi, n_0, d \cdot \pi_0),$$

where $\forall c > 0$, $\forall d > 0$, $d \cdot \pi \leq 1$, $d \cdot \pi_0 \leq 1$. Additional design consideration includes interpretability and being bounded within $(0, 1)$.

Based on the criteria discussed above, we propose the following loss function:

$$L(n, \pi, n_0, \pi_0) = \sigma\left(w\frac{n}{n_0} + (1-w)\frac{\pi_0}{\pi} - 1\right),$$

where w is a weighing parameter, and $\sigma(\cdot)$ is a link function defined as $\sigma(x) = \frac{1}{1+e^{-x}}$. The link function $\sigma(\cdot)$ scales the loss score to the range of $(0, 1)$. The parameter w determines the trade-off between reducing sample size and increasing power. We performed a sensitivity analysis on w and found that the sample size and power were invariant to w when $w > 0.4$ (Supplementary Figure S1). Within the range of $w \leq 0.4$, a larger w gives greater priority to reducing the sample size, while a smaller w prioritizes increasing power. Therefore, we recommend using $w = 0.5$, which assigns equal importance to sample size reduction and power improvement and provides an optimized balance between sample size and power. We recommend calculating n_0 using the standard sample size formula for testing $H_0 : p_E - p_C = 0$ and $H_a : p_E - p_C > 0$ [2]. When the sample size is smaller than that of the standard two-sample test and the power is greater, which is the most desirable scenario, the sum of the first two components is smaller than 1, resulting in a loss score smaller than 0.5; when the sample size and power match those of the standard two-sample test, the loss score equals 0.5; when the sample size is larger and the power is lower, the loss score is greater than 0.5, which is considered undesirable.

The optimal parameters for the inconclusive regions are selected as the smallest solution set that minimizes the loss function. Firstly, given minimum target power $\pi_{\min}$ and sample size n, we specify a pair of $(\gamma_{\max}, \lambda_{\max})$. Then the optimal design sample size and the corresponding power (N^*, π^*) are determined by minimizing the loss score. Formally,

$$\left(N^*, \pi^*\right) = \arg\min_{N \in Q} L(N, \pi, n_0, \pi_0 | \gamma_{\max}, \lambda_{\max}), \qquad (2.3)$$

where Q represents the search set for the sample size. An example illustrating the determination of $\gamma_{\max}$ and $\lambda_{\max}$ is provided in Table S1 in the Supplementary Materials. In practice, we impose an additional constraint on power π to ensure that it remains at a relatively high level, requiring $\pi \geq \pi_{\min} - c$, where c is a constant (e.g., $c = 0.05$). When comparing different trial configurations, the loss score provides a systematic evaluation of both sample size and power.

Table 1. Optimal design parameters of the TDR one-stage 2-by-2 design with $\alpha_{\max} = 0.20$, $\beta_{\max} = 0.20$, $\pi_{\min} = 0.80$, *and* $c = 0.05$.

Setting			Design Parameter										
δ	p_C	p_E	$\gamma_{\max}$	$\lambda_{\max}$	s	m	N	π	β	α	γ	η	λ
0.15	0.10	0.25	0.08	0.20	1	4	44	0.79	0.13	0.15	0.08	0.25	0.16
0.20	0.10	0.30	0.08	0.20	1	3	28	0.80	0.13	0.14	0.08	0.23	0.15
0.25	0.10	0.35	0.12	0.20	1	3	22	0.77	0.11	0.08	0.11	0.27	0.19
0.15	0.20	0.35	0.16	0.30	0	8	54	0.76	0.08	0.15	0.16	0.42	0.29
0.20	0.20	0.40	0.16	0.30	0	5	30	0.77	0.08	0.16	0.16	0.43	0.30
0.25	0.20	0.45	0.10	0.15	1	4	24	0.81	0.13	0.17	0.06	0.22	0.14
0.15	0.30	0.45	0.11	0.20	1	12	62	0.76	0.14	0.17	0.10	0.28	0.19
0.20	0.30	0.50	0.07	0.20	1	8	40	0.81	0.12	0.19	0.06	0.24	0.15
0.15	0.35	0.50	0.10	0.20	1	13	60	0.76	0.15	0.19	0.09	0.26	0.17
0.20	0.35	0.55	0.10	0.15	1	9	40	0.81	0.13	0.20	0.06	0.23	0.15
0.25	0.35	0.60	0.10	0.20	1	6	24	0.78	0.15	0.18	0.08	0.24	0.16
0.20	0.40	0.60	0.10	0.20	1	10	38	0.76	0.14	0.16	0.10	0.27	0.19
0.25	0.40	0.65	0.16	0.30	0	7	24	0.77	0.07	0.15	0.16	0.43	0.29
0.15	0.50	0.65	0.10	0.15	2	20	70	0.77	0.18	0.19	0.05	0.17	0.11
0.20	0.50	0.70	0.11	0.20	1	12	38	0.77	0.13	0.16	0.10	0.28	0.19
0.15	0.55	0.70	0.13	0.25	0	18	56	0.78	0.09	0.20	0.13	0.36	0.24
0.20	0.55	0.75	0.14	0.30	0	11	32	0.79	0.08	0.19	0.13	0.38	0.26
0.25	0.60	0.85	0.10	0.20	1	7	18	0.77	0.17	0.19	0.06	0.22	0.14
0.20	0.65	0.85	0.10	0.20	1	11	28	0.78	0.15	0.18	0.07	0.24	0.15
0.15	0.70	0.85	0.11	0.25	1	19	48	0.79	0.14	0.19	0.07	0.24	0.16

$\gamma_{\max}$: design constraint for γ; $\lambda_{\max}$: design constraint for λ; s: statistical difference boundary; m: clinical relevance boundary; N: total sample size; π: power; α: type I error; β: type II error; γ: inconclusive probability under H_a; η: inconclusive probability under H_0; λ: average inconclusive probability under H_0 and H_a.

3. NUMERICAL STUDIES

3.1 TDR One-Stage Design

Table 1 lists the optimal TDR one-stage 2-by-2 design parameters with varying differences in response rate, $\delta = p_E - p_C \in \{0.15, 0.20, 0.25\}$, under target levels $\alpha_{\max} = 0.20$, $\beta_{\max} = 0.20$, $\pi_{\min} = 0.80$, and $c = 0.05$. In this table, $\gamma_{\max}$ and $\lambda_{\max}$ are design parameters corresponding to the optimal sample size, determined using the loss function with a weight parameter $w = 0.50$, which equally weighs the importance of sample size and power. For each case of response rates, we employ a 1:1 randomization. The total sample size required for the trial is denoted by N with corresponding decision boundaries s and m. To evaluate the proposed design, we compare the operating characteristics of the TDR design with the method proposed by [6] (HW) and the method proposed by [10] (LBR). The comparison evaluates the percentage reduction in sample size relative to the conventional calculation for two-sample proportions under the same settings of type I error $\alpha_{\max}$, and type II error $\beta_{\max}$. The results are shown in Figure 1. Overall, the TDR design provides 26.7–51.7% sample size savings as compared to the conventional approach, while the HW method provides up to 22.7% sample size savings and the LBR method provides 20.0–42.6% sample size savings. In most of the scenarios, the proposed method outperforms the HW method and the LBR method with minimal loss of power. The TDR method generally yields higher power than the HW method and provides more sample size savings in all cases. Compared to the LBR method, the TDR method provides superior or comparable sample size reduction up to 13.8% except in one case. TDR achieved 0.3–12.3% type II error reduction at the cost of up to 6.3% power loss and gained 1.0% power in one case. In terms of the type I error, all three methods are comparable and are constrained below 0.20. In the one case where the TDR method does not show sample size saving compared to the LBR method, the type II error is 7.5% lower and the power is 1.1% higher at response rates of (0.30, 0.50). In four of the twenty cases, the HW method does not exhibit substantial sample size saving and is therefore not displayed.

Under more stringent requirements on type I and type II errors, i.e., $\alpha_{\max} = 0.10$, $\beta_{\max}0.10$, $\pi_{\min} = 0.80$, and $c = 0.05$, superior performances of the proposed method are observed in most cases. Table 2 provides details of the optimal trial design parameters. Comparisons of sample size reduction, as well as operating characteristics, are shown in Figure 2. Overall, the TDR design achieves 37.2–57.0% sample size reduction, the HW design achieves up to 34.0% sample size reduction, and the LBR design achieves 37.2–49.1% sample size reduction. The TDR method provides additional 2.9–18.0% sample size savings as compared to the LBR method in 18 of the 20 cases. In two cases where equal sample sizes are calculated, TDR provides 3.1% and 1.2%

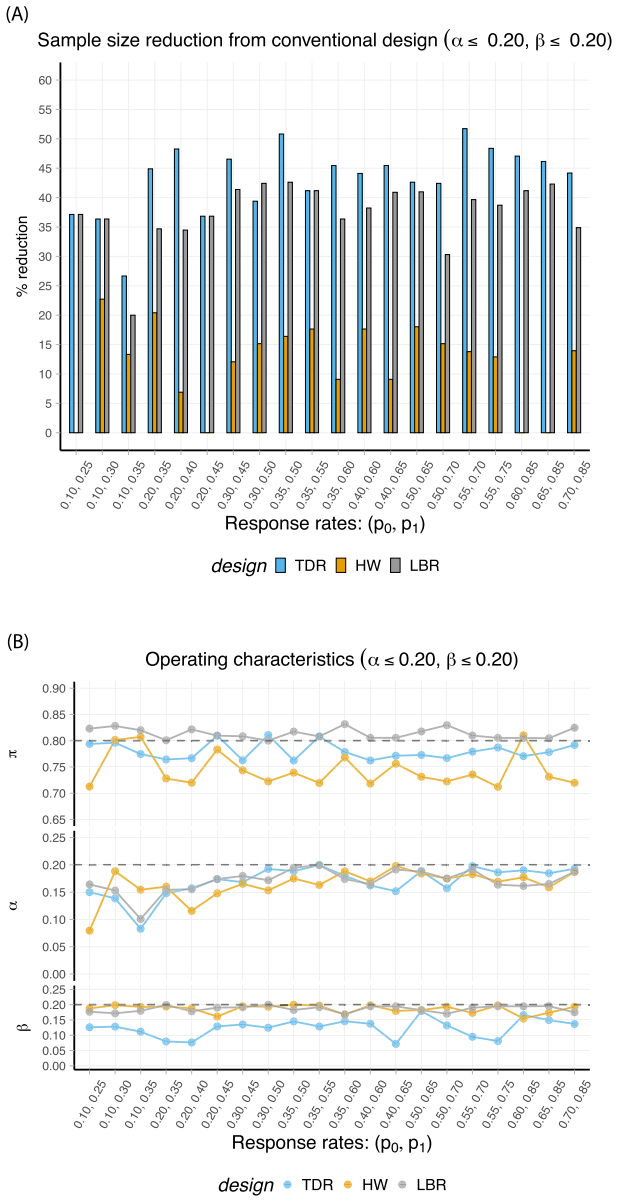

Figure 1: Comparison of TDR one-stage 2-by-2 with the HW method [6] and LBR method [10] under $\alpha_{\max} = 0.20$, $\beta_{\max} = 0.20$, $\pi_{\min} = 0.80$, and $c = 0.05$. (A) Sample size reduction with respect to the conventional sample size calculation for two-sample proportions; (B) Comparison of operating characteristics power π, type I error α, and type II error β.

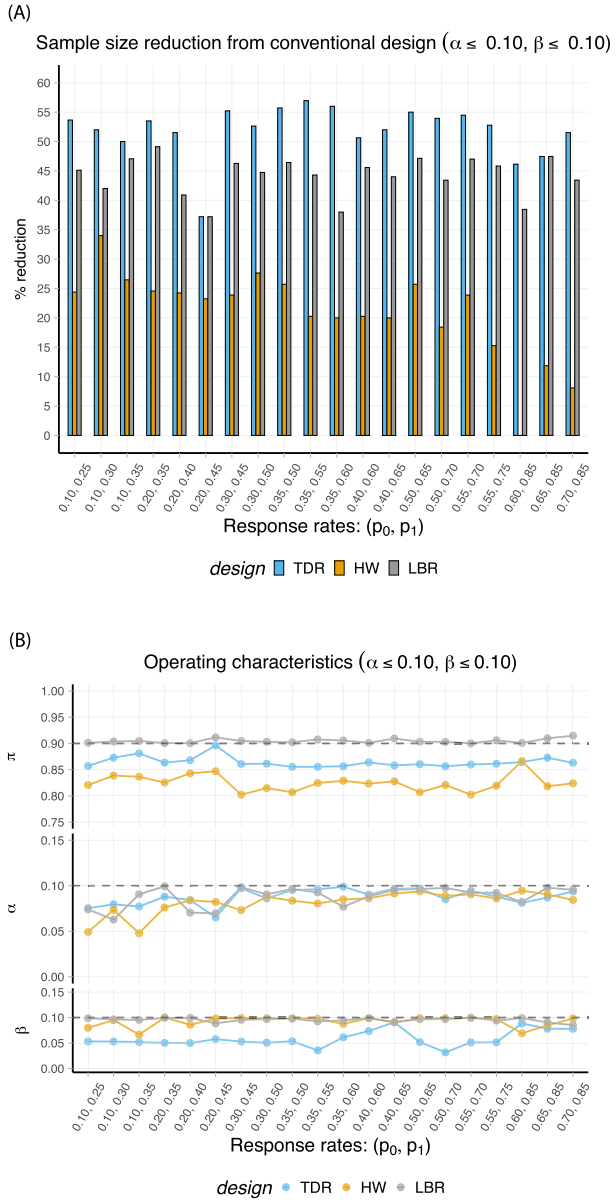

Figure 2: Comparison of TDR one-stage 2-by-2 with the HW method [6] and LBR method [10] under $\alpha_{\max} = 0.10$, $\beta_{\max} = 0.10$, $\pi_{\min} = 0.90$, and $c = 0.05$. (A) Sample size reduction with respect to the conventional sample size calculation for two-sample proportions; (B) Comparison of operating characteristics power π, type I error α, and type II error β.

reduction in type II errors at response rates of (0.20, 0.45) and (0.65, 0.85), respectively.

3.2 TDR Two-Stage Design

Extending the method to a two-stage design, we provide details of the TDR two-stage 2-by-2 design in Table 3. In addition, given the established sample size saving performance of the LBR method proposed by [10], we use it as a reference to benchmark the proposed method in terms of expected sample size (EN) and maximum sample size as shown in Figure 3. The total sample sizes required for stage 1 and stage 2 of the trial are denoted by N_1 and N_2, respectively, with corresponding decision boundaries s_1 and m_1 for stage 1, and s_2 and m_2 for stage 2.

Table 2. Optimal design parameters of the TDR one-stage 2-by-2 design with $\alpha_{\max} = 0.10$, $\beta_{\max} = 0.10$, $\pi_{\min} = 0.90$, *and* $c = 0.05$.

Setting			Design Parameter										
δ	p_C	p_E	$\gamma_{\max}$	$\lambda_{\max}$	s	m	N	π	β	α	γ	η	λ
0.15	0.10	0.25	0.10	0.25	1	7	76	0.86	0.05	0.08	0.09	0.35	0.22
0.20	0.10	0.30	0.10	0.20	1	5	48	0.87	0.05	0.08	0.07	0.32	0.20
0.25	0.10	0.35	0.10	0.20	1	4	34	0.88	0.05	0.08	0.07	0.31	0.19
0.15	0.20	0.35	0.10	0.25	1	15	106	0.86	0.05	0.09	0.09	0.36	0.23
0.20	0.20	0.40	0.11	0.25	1	10	64	0.87	0.05	0.08	0.08	0.35	0.22
0.25	0.20	0.45	0.10	0.15	2	9	54	0.90	0.06	0.07	0.05	0.24	0.14
0.15	0.30	0.45	0.10	0.25	1	23	120	0.86	0.05	0.10	0.09	0.36	0.22
0.20	0.30	0.50	0.10	0.25	1	15	72	0.86	0.05	0.09	0.09	0.36	0.23
0.15	0.35	0.50	0.10	0.25	1	27	124	0.86	0.05	0.10	0.09	0.37	0.23
0.20	0.35	0.55	0.11	0.30	0	16	68	0.86	0.04	0.10	0.11	0.45	0.28
0.25	0.35	0.60	0.10	0.25	1	11	44	0.86	0.06	0.10	0.08	0.34	0.21
0.20	0.40	0.60	0.10	0.20	2	20	78	0.86	0.07	0.09	0.06	0.27	0.17
0.25	0.40	0.65	0.10	0.15	2	13	48	0.86	0.09	0.10	0.05	0.23	0.14
0.15	0.50	0.65	0.10	0.25	1	37	126	0.86	0.05	0.10	0.09	0.37	0.23
0.20	0.50	0.70	0.12	0.30	0	22	70	0.86	0.03	0.09	0.11	0.46	0.29
0.15	0.55	0.70	0.12	0.30	1	39	122	0.86	0.05	0.09	0.09	0.37	0.23
0.20	0.55	0.75	0.10	0.25	1	23	68	0.86	0.05	0.09	0.09	0.36	0.23
0.25	0.60	0.85	0.10	0.20	2	16	42	0.86	0.09	0.08	0.05	0.24	0.14
0.20	0.65	0.85	0.10	0.20	2	24	62	0.87	0.08	0.09	0.05	0.26	0.15
0.15	0.70	0.85	0.10	0.25	2	38	96	0.86	0.08	0.09	0.06	0.28	0.17

$\gamma_{\max}$: design constraint for γ; $\lambda_{\max}$: design constraint for λ; s: statistical difference boundary; m: clinical relevance boundary; N: total sample size; π: power; α: type I error; β: type II error; γ: inconclusive probability under H_a; η: inconclusive probability under H_0; λ: average inconclusive probability under H_0 and H_a.

Table 3. Optimal design parameters of the TDR two-stage 2-by-2 design with $\alpha_{\max} = 0.20$, $\beta_{\max} = 0.20$, $\pi_{\min} = 0.80$, *and* $c = 0.05$.

Setting		Design Parameter														
p_C	p_E	$\gamma_{\max}$	$\lambda_{\max}$	s_1	m_1	s_2	m_2	EN	N_1	N_2	π	β	α	γ	η	λ
0.10	0.25	0.10	0.07	−4	3	1	4	47.63	46	50	0.85	0.07	0.19	0.03	0.09	0.06
0.10	0.30	0.10	0.07	−3	2	1	3	29.65	28	32	0.85	0.07	0.17	0.03	0.11	0.07
0.10	0.35	0.10	0.07	−2	2	1	3	25.34	24	28	0.88	0.05	0.14	0.03	0.11	0.07
0.20	0.35	0.06	0.15	−4	6	1	9	60.90	56	66	0.82	0.13	0.17	0.06	0.16	0.11
0.20	0.40	0.10	0.10	−4	4	1	6	36.69	34	40	0.83	0.15	0.17	0.05	0.14	0.09
0.20	0.45	0.10	0.14	−3	3	1	5	25.75	24	28	0.80	0.08	0.12	0.08	0.18	0.13
0.30	0.45	0.06	0.15	−4	10	1	13	66.93	64	70	0.81	0.14	0.19	0.06	0.15	0.10
0.30	0.50	0.10	0.09	−5	6	1	8	37.86	36	40	0.81	0.09	0.19	0.05	0.13	0.09
0.35	0.50	0.10	0.13	−7	11	1	14	61.97	60	64	0.76	0.10	0.17	0.09	0.17	0.13
0.35	0.55	0.10	0.16	−3	7	1	10	39.98	38	42	0.77	0.09	0.14	0.09	0.20	0.15
0.35	0.60	0.10	0.15	−2	5	1	8	28.78	26	32	0.82	0.14	0.14	0.07	0.18	0.13
0.40	0.60	0.10	0.14	−3	8	1	11	39.96	38	42	0.78	0.09	0.15	0.09	0.19	0.14
0.40	0.65	0.10	0.09	−4	6	1	8	27.70	26	30	0.83	0.08	0.18	0.04	0.11	0.08
0.50	0.65	0.10	0.15	−6	15	1	21	64.97	58	72	0.79	0.14	0.18	0.07	0.16	0.11
0.50	0.70	0.16	0.25	−3	9	1	12	35.93	34	38	0.77	0.18	0.16	0.09	0.18	0.13
0.55	0.70	0.10	0.11	−6	16	1	19	57.94	56	60	0.78	0.11	0.20	0.06	0.14	0.10
0.55	0.75	0.15	0.20	−3	10	1	13	35.84	34	38	0.78	0.17	0.15	0.08	0.17	0.13
0.60	0.85	0.10	0.08	−2	6	1	9	20.78	18	24	0.84	0.09	0.19	0.03	0.12	0.08
0.65	0.85	0.15	0.25	−3	9	0	12	27.97	26	30	0.80	0.11	0.17	0.11	0.24	0.18
0.70	0.85	0.10	0.15	−3	16	1	19	45.91	44	48	0.79	0.19	0.19	0.06	0.14	0.10

$\gamma_{\max}$: design constraint for γ; $\lambda_{\max}$: design constraint for λ; s_1: statistical difference boundary in stage 1; m_1: clinical relevance boundary in stage 1; s_2: statistical difference boundary in stage 2; m_2: clinical relevance boundary in stage 2; EN: expected total sample size; N_1: total sample size in stage 1; N_2: total sample size in stage 2; π: power; α: type I error; β: type II error; γ: inconclusive probability under H_a; η: inconclusive probability under H_0; λ: average inconclusive probability under H_0 and H_a.

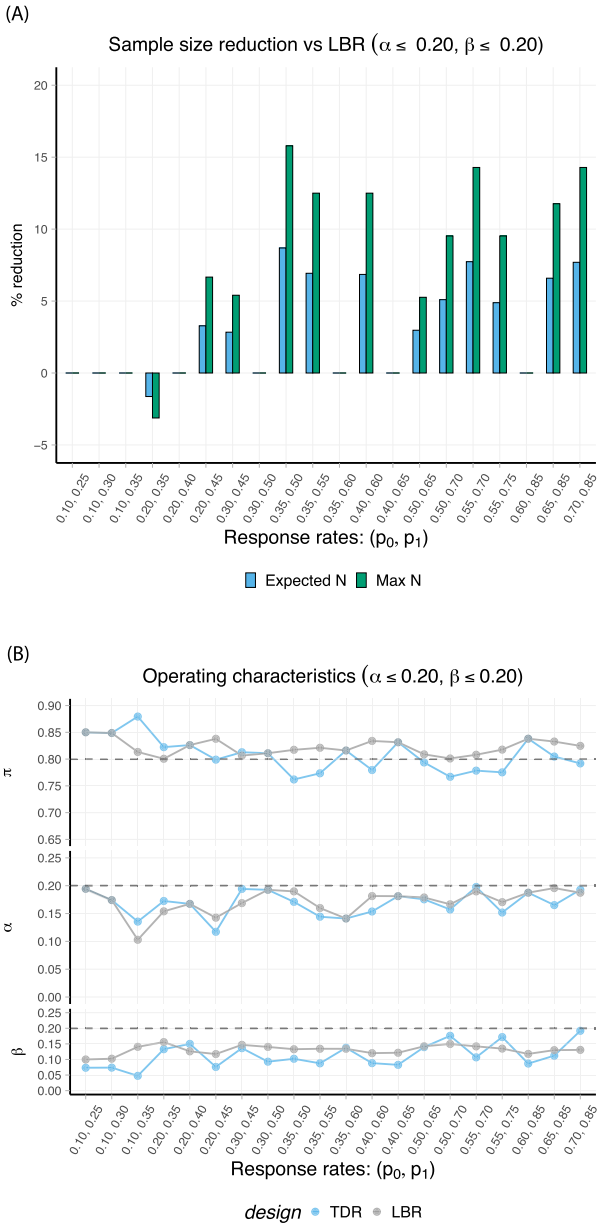

Figure 3: Comparison of TDR two-stage 2-by-2 with the LBR method [10] under $\alpha_{\max} = 0.20$, $\beta_{\max} = 0.20$, $\pi_{\min} = 0.80$, and $c = 0.05$. (A) Expected sample size reduction and maximum sample size reduction with respect to the LBR method; (B) Comparison of operating characteristics power π, type I error α, and type II error β.

In a two-stage TDR design, the inconclusive region is considered only in stage 2. As a result, if the sample size in stage 1 is much larger than in stage 2, the potential for sample size savings will be limited. In 11 of the 20 cases, our proposed method provides a 2.8–8.7% reduction in expected sample size and a 5.4–15.8% reduction in maximum sample size as compared to the LBR method. In all cases, the proposed method shows reductions in type II errors compared to the LBR method. In the nine cases of no sample size reduction, we observe reductions in type II errors except for minimal type II error inflation in two cases (0.02 and 0.003 at the response rates (0.20, 0.40) and (0.35 and 0.60), respectively). This could be due to over-constraining design parameters, which could be remedied by a more granular search of $\gamma_{\max}$ and $\lambda_{\max}$ in regions around the current constraints, adjusting the loss function weight parameter w, or inspecting the sculpting boundaries. For example, at $p_C = 0.35$, $p_E = 0.60$, by decreasing the statistical difference boundary, s_2, one could reduce the total sample size by two in the second stage with a 1.8% reduction in type I error with a power still higher than 0.75.

3.3 Sensitivity Analysis

As previously shown, the TDR design can be applied to more stringent requirements on type I and type II errors. To account for the effect of variation in control response rates, we apply the type I error constraints to the maximum of type I errors yielded from a confidence interval of p_C. A confidence interval of 30% is chosen for demonstration purposes, as with historical data, there could be a fairly informed estimation of control response rates. In general, the proposed design shows superior performance to the HW design, the LBR design, as well as the conventional design. The details can be found in Table S2 and Figure S2 of the Supplementary Materials.

We also conducted a comparison between the proposed 2-by-2 design and the 3-by-2 design. Further details can be found in Table S3 and Figure S3 in the Supplementary Materials. Overall, with more granular control of the inconclusive region, the 3-by-2 TDR design provides a reduction in type II error, which is an advantage of using the 3-by-2 design. Depending on true response rates, the 3-by-2 design may provide sample size savings in some cases. However, one should also take note of the increased design complexity when choosing between 2-by-2 and 3-by-2 designs.

4. TRIAL APPLICATION

Defachelles et al. [3] conducted a randomized two-parallel group phase II trial to evaluate the efficacy and safety of the vincristine-irinotecan combination with and without temozolomide (VIT and VI, respectively) among patients with relapsed or refractory rhabdomyosarcoma. In this study, a total of 120 patients were randomized 1:1 to receive 21-day cycles of VI or VIT, with 60 patients per arm. The primary endpoint is the objective response rate (ORR) after two cycles. Originally designed as a non-comparative randomized phase II trial, the trial performed Simon's two-stage design [19] in each arm to define the sample size. The design parameters are set as $p_0 = 0.35$ under the null hypothesis and $p_1 = 0.55$ under the alternative hypothesis for each arm. A dropout rate of 8% was considered in this trial. The ORR after two cycles in the whole population was 44% in the VIT arm and 31% in the VI arm (i.e., $\hat{p}_E = 0.44$ and $\hat{p}_C = 0.31$).

Table 4. Application of the TDR two-stage design and the LBR method to the VIT-0910 trial.

	N	α	β	π
VIT-0910	128	0.10	0.10	0.90
TDR	102	0.09	0.02	0.90
LBR	102	0.09	0.05	0.90

N: total sample size; π: power; α: type I error; β: type II error. An 8% dropout rate is considered in the total sample size.

The TDR two-stage design, as detailed in Section 2.2, is performed to re-calculate the sample size in the VIT-0910 trial. In adherence to the above trial configurations, we set $p_C = 0.35$ and $p_E = 0.55$ in our design. We search for the sample size under a specified type I error of $\alpha_{\max} = 0.10$ and seek to achieve a power of 0.90. The selection of the optimal sample size is based on minimizing the loss score across all potential candidates. As a comparison, we also compute the sample size using the LBR design. The summarized sample sizes and design parameters are shown in Table 4. Both our TDR design and the LBR design exhibit a substantial decrease in the required sample size for the same trial.

5. DISCUSSIONS

In this paper, we propose a three-outcome dual-criterion randomized phase II design that utilizes inconclusive region sculpting to reduce sample size and type II error. The proposed TDR trial design shows sample size saving and reduction in type II error compared to existing methods. When the requirements for type I and type II error control become more stringent, such as controlling α and β to be within 10% instead of 20%, the proposed method demonstrates even greater sample size savings. While the benefit of sample savings and type II error reduction is evident in most cases, a limitation of the proposed design is a slight reduction in power. However, this can be controlled by specifying an acceptable power threshold and adjusting design parameters accordingly. It should also be noted that as the trials of interest for this design consider binary outcomes, the discreteness of responses may lead to fluctuations in type I and type II errors, as well as power, when automatically searching design parameters over a range of values for $(\gamma_{\max}, \lambda_{\max})$. This can be mitigated by manually adjusting the sample size or design parameters, and the loss function can assist in such an adjustment process by systematically evaluating the trade-off between power and sample size.

The TDR design provides flexibility for an extension to a two-stage setting, particularly when early stopping due to lack of efficacy is an ethical consideration. Additionally, the inconclusive region can be more finely sculpted using a 3-by-2 TDR design, further reducing type II error. To align with specific study objectives, parameters and loss function settings can be adjusted to control type I and type II errors, inconclusive probabilities, randomization ratio, confidence interval of p_C for robustness, and early stopping probabilities. Moreover, given the flexibility of the design, the dual criteria on clinical relevance could potentially be extended to a three-region decision framework to further control the inconclusive probabilities.

SUPPLEMENTARY MATERIAL

Supplementary Material for Three-Outcome Dual-Criterion Randomized Phase II Clinical Trial Design.

ACKNOWLEDGEMENTS

We would like to express our gratitude to the Editor, the Associate Editor, and two reviewers for their valuable comments and suggestions, which significantly contributed to improving the quality of the article.

FUNDING

Lin's research is partially supported by NIH/NCI grants R01CA261978 and 1R21LM014699.

Accepted 3 March 2025

REFERENCES

[1] Brookmeyer, R. and Crowley, J. (1982). A confidence interval for the median survival time. *Biometrics* **38**(1) 29–41. https://doi.org/10.2307/2530286

[2] Chow, S.-C., Wang, H. and Shao, J. (2007). *Sample Size Calculations in Clinical Research*, 2nd edn. Chapman and Hall/CRC. https://doi.org/10.1201/9781584889830. MR2356591

[3] Defachelles, A.-S., Bogart, E., Casanova, M., Merks, J. H. M., Bisogno, G., Calareso, G., Gallego Melcon, S., Gatz, S. A., Le Deley, M.-C., McHugh, K., Probst, A., Rocourt, N., van Rijn, R. R., Wheatley, K., Minard-Colin, V. and Chisholm, J. C. (2021). Randomized phase II trial of vincristine-irinotecan with or without temozolomide, in children and adults with relapsed or refractory rhabdomyosarcoma: a European paediatric soft tissue sarcoma study group and innovative therapies for children with cancer trial. *Journal of Clinical Oncology* **39**(27) 2979–2990. https://doi.org/10.1200/JCO.21.00124.

[4] Fisch, R., Jones, I., Jones, J., Kerman, J., Rosenkranz, G. K. and Schmidli, H. (2015). Bayesian design of proof-of-concept trials. *Therapeutic Innovation & Regulatory Science* **49**(1) 155–162. https://doi.org/10.1177/2168479014533970.

[5] Herson, J. and Carter, S. K. (1986). Calibrated phase II clinical trials in oncology. *Statistics in Medicine* **5**(5) 441–447. https://doi.org/10.1002/sim.4780050508.

[6] Hong, S. and Wang, Y. (2007). A three-outcome design for randomized comparative phase II clinical trials. *Statistics in Medicine* **26**(19) 3525–3534. https://doi.org/10.1002/sim.2824. MR2393733

[7] Kola, I. and Landis, J. (2004). Can the pharmaceutical industry reduce attrition rates? *Nature Reviews. Drug Discovery* **3**(8) 711–715. https://doi.org/10.1038/nrd1470.

[8] Law, M., Grayling, M. J. and Mander, A. P. (2021). A stochastically curtailed two-arm randomised phase II trial design for binary outcomes. *Pharmaceutical Statistics* **20**(2) 212–228. https://doi.org/10.1002/PST.2067.

[9] Lee, J. J. and Feng, L. (2005). Randomized phase II designs in cancer clinical trials: current status and future directions. *Journal of Clinical Oncology: Official Journal of the American Society of Clinical Oncology* **23**(19) 4450–4457. https://doi.org/10.1200/JCO.2005.03.197.

[10] Litwin, S., Basickes, S. and Ross, E. A. (2017). Two-sample binary phase 2 trials with low type I error and low sample size. *Statistics in Medicine* **36**(9) 1383–1394. https://doi.org/10.1002/sim.7226. MR3631967

[11] Mozgunov, P., Jaki, T. and Gasparini, M. (2019). Loss functions in restricted parameter spaces and their Bayesian applications. *Journal of Applied Statistics* **46**(13) 2314. https://doi.org/10.1080/02664763.2019.1586848. MR3987561

[12] Mozgunov, P. and Jaki, T. (2019). An information theoretic phase I–II design for molecularly targeted agents that does not require an assumption of monotonicity. *Journal of the Royal Statistical Society. Series C, Applied Statistics* **68**(2) 347. https://doi.org/10.1111/rssc.12293. MR3902998

[13] Rubinstein, L., Crowley, J., Ivy, P., LeBlanc, M. and Sargent, D. (2009). Randomized phase II designs. *Clinical Cancer Research* **15**(6) 1883–1890. https://doi.org/10.1158/1078-0432.CCR-08-2031.

[14] Rubinstein, L. V., Korn, E. L., Freidlin, B., Hunsberger, S., Ivy, S. P. and Smith, M. A. (2005). Design issues of randomized phase ii trials and a proposal for phase II screening trials. *Journal of Clinical Oncology* **23**(28) 7199–7206. https://doi.org/10.1200/JCO.2005.01.149.

[15] Sargent, D. J., Chan, V. and Goldberg, R. M. (2001). A three-outcome design for phase II clinical trials. *Controlled Clinical Trials* **22**(2) 117–125. https://doi.org/10.1016/S0197-2456(00)00115-X.

[16] Shan, M. (2021). A confidence function-based posterior probability design for phase II cancer trials. *Pharmaceutical Statistics* **20**(3) 485–498. https://doi.org/10.1002/pst.2089.

[17] Sharma, M. R., Stadler, W. M. and Ratain, M. J. (2011). Randomized phase II trials: a long-term investment with promising returns. *JNCI Journal of the National Cancer Institute* **103**(14) 1093–1100. https://doi.org/10.1093/jnci/djr218.

[18] Simon, R., Wittes, R. E. and Ellenberg, S. S. (1985). Randomized phase II clinical trials. *Cancer Treatment Reports* **69**(12) 1375–1381.

[19] Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* **10**(1) 1–10. https://doi.org/10.1016/0197-2456(89)90015-9. MR4366283

[20] Storer, B. E. (1992). A class of phase II designs with three possible outcomes. *Biometrics* **48**(1) 55–60. https://doi.org/10.2307/2532738.

[21] Wouters, O. J., McKee, M. and Luyten, J. (2020). Estimated research and development investment needed to bring a new medicine to market, 2009–2018. *JAMA* **323**(9) 844–853. https://doi.org/10.1001/jama.2020.1166.

Yujia Wang. Department of Bioinformatics and Computational Biology, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.
E-mail address: ywang74@mdanderson.org

Xiaohan Chi. Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.
E-mail address: xchi@mdanderson.org

Ruitao Lin. Department of Biostatistics, The University of Texas MD Anderson Cancer Center, Houston, Texas 77030, USA.
E-mail address: rlin@mdanderson.org