Received: 14 October 2018 | Revised: 16 April 2019 | Accepted: 22 April 2019

DOI: 10.1002/bimj.201800320

Biometrical Journal

RESEARCH PAPER

On construction of single-arm two-stage designs with consideration of both response and toxicity

Huan Yin[1] | **Weizhen Wang**[1,2] | **Zhongzhan Zhang**[1]

[1]College of Applied Sciences, Beijing University of Technology, Beijing, P. R. China

[2]Department of Mathematics and Statistics, Wright State University, Dayton, OH, USA

Correspondence
Weizhen Wang, Department of Mathematics and Statistics, Wright State University, Dayton, OH 45435, USA.
Email: weizhen.wang@wright.edu

Funding information
National Natural Science Foundation of China, Grant/Award Number: 11271039; Education Ministry Funds, Grant/Award Number: 006000546615539

Abstract

When establishing a treatment in clinical trials, it is important to evaluate both effectiveness and toxicity. In phase II clinical trials, multinomial data are collected in m-stage designs, especially in two-stage ($m = 2$) design. Exact tests on two proportions, p_r for the response rate and p_t for the nontoxicity rate, should be employed due to limited sample sizes. However, existing tests use certain parameter configurations at the boundary of null hypothesis space to determine rejection regions without showing that the maximum Type I error rate is achieved at the boundary of null hypothesis. In this paper, we show that the power function for each test in a large family of tests is nondecreasing in both p_r and p_t; identify the parameter configurations at which the maximum Type I error rate and the minimum power are achieved and derive level-α tests; provide optimal two-stage designs with the least expected total sample size and the optimization algorithm; and extend the results to the case of $m > 2$. Some R-codes are given in the Supporting Information.

KEYWORDS

conditional distribution, maximum Type I error rate, minimum power, multinomial distributions, phase II clinical trials

1 | INTRODUCTION

To assess a treatment in a phase II clinical trial for cancer study, a one-arm study is typically employed. That is, all patients receive the treatment. From each patient, we observe the outcomes of both response and nontoxicity. Suppose a binary variable $Z_{ij\upsilon}$ is observed from the υth patient for $\upsilon = 1, 2, \ldots$ and $i, j = 0$ or 1. Here, i and j are the indicators of whether the υth patient shows improvement and nontoxicity, respectively, after receiving the treatment. For example, $Z_{10\upsilon}$ is an indicator function for $(i, j) = (1, 0)$, and $\{Z_{10\upsilon} = 1\}$ means that the υth patient shows improvement and toxicity. The υth patient has four possible outcomes as follows:

		Nontoxicity (j)	
		Yes (1)	**No (0)**
Response (i)	Yes (1)	$Z_{11\upsilon}, p_{11}$	$Z_{10\upsilon}, p_{10}$
	No (0)	$Z_{01\upsilon}, p_{01}$	$Z_{00\upsilon}, p_{00}$

where $p_{ij} = P(Z_{ij\upsilon} = 1)$ is assumed unchanged in υ. Thus, $(Z_{11\upsilon}, Z_{10\upsilon}, Z_{01\upsilon}, Z_{00\upsilon})$ follows a multinomial distribution with only one trial and probabilities p_{ij}s. Let $p_r = p_{11} + p_{10}$ denote the improvement rate and $p_t = p_{11} + p_{01}$ denote the nontoxicity

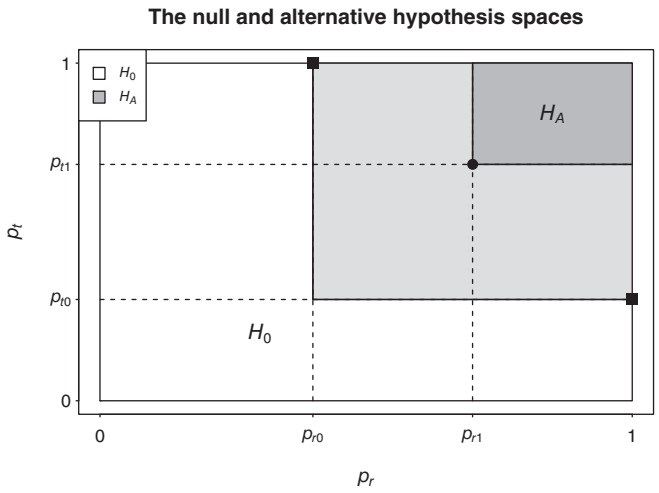

FIGURE 1 The hypotheses, H_0 (the white area) and H_A (the dark gray area) in (1), two parameter points $\underline{p}^*_{0r}$ (the higher black square) and $\underline{p}^*_{0t}$ (the lower black square) in H_0 at which the maximum Type I error rate is achieved, and one parameter point $\underline{p}^*_A$ (the black circle) at which the minimum power is achieved

rate. Here we consider the nontoxicity rate for simplicity. The first goal of this article is to derive exact tests based on Z_{ijv}s for the hypotheses:

$$H_0 : p_r \le p_{r_0} \text{ or } p_t \le p_{t_0} \text{ versus } H_A : p_r \ge p_{r_1} \text{ and } p_t \ge p_{t_1}, \tag{1}$$

where p_{r_0}, p_{t_0}, p_{r_1}, and p_{t_1} are four prespecified proportions satisfying $p_{r_0} < p_{r_1}$ and $p_{t_0} < p_{t_1}$ (see Figure 1 for the hypotheses), H_0 represents either a noneffective or unsafe treatment which should be screened out from further testing, and H_A represents an effective and safe treatment to warrant further testing, for example, in phase III clinical trials. Different from the traditional theory of hypothesis testing, H_A here is a selected subset of the complement of H_0 for which we want to guarantee sufficient power. Due to ethical and economic considerations, an m-stage ($m \ge 2$) design is often used because it gives the benefits of sample size saving and patient protection by ending the trial earlier. When $m = 2$, it is a two-stage design. The second goal of this article is to determine the number of patients in each stage that minimizes the expected total sample size for a given upper limit for the total number of possible patients.

Example 1.1. Suppose we can recruit up to $N^* = 65$ patients to test (1) for $(p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1}) = (0.4, 0.4, 0.6, 0.6)$. The test level α is set to be 0.05, and the minimum power $1 - \beta$ is at least 0.8. We can select up to two groups of patients sequentially (in Stages 1 and 2) of sizes n_1 and n_2 ($n_1 + n_2 \le N^*$). Depending on the outcomes Z_{ijv}s in the first group, we decide whether a second group of patients is selected in Stage 2, that is, the trial ends randomly in Stage 1 or 2 with a total sample size n_1 or $n_1 + n_2$. The following two questions are of great interest to derive an optimal two-stage design:

(a) For a given pair of (n_1, n_2), how do we construct a rejection region R with the predetermined test level α and minimum power $1 - \beta$?
(b) Which pair of (n_1, n_2) provides the least expected total sample size under H_0 or H_A?

The answers will be given in Section 3.

The hypotheses (1) are a natural extension of

$$H_0 : p_r \le p_{r_0} \text{ versus } H_A : p_r \ge p_{r_1}, \tag{2}$$

which only addresses the effectiveness of a treatment, leaving the safety out of consideration. The observation Z_{ijv} reduces to Z_{iv}, which only records the status on effectiveness and ignores the outcome on toxicity for each patient. Simon (1989) derived optimal two-stage designs for this case. Shan et al. (2016b) proved that Simon's designs controlled the maximum Type I error rate. Furthermore, Banerjee and Tsiatis (2006), Lin and Shih (2004), Jin and Wei (2012), and Shan et al. (2016a) proposed adaptive two-stage designs, in which the sample size n_2 in Stage 2 depends on the outcome in Stage 1. These designs are certainly attractive to practitioners.

In the past two decades, the phase II clinical trials taking both treatment response and safety as primary endpoints have been investigated in literature. Thall, Simon, and Estey (1995), Thall and Sung (1998), and Stallard, Thall, and Whitehead (1999) proposed Bayesian methods. A drawback of using Bayesian methods is that if the prior information is incorrect, then the

TABLE 1 The contingency tables for response and nontoxicity in Stages 1 and 2 of a two-stage design

			Nontoxicity		
			Yes (1)	**No (0)**	**Total**
Stage 1	Response	Yes(1)	X_{11}, p_{11}	X_{10}, p_{10}	X_r, p_r
		No (0)	X_{01}, p_{01}	X_{00}, p_{00}	
	Total		X_t, p_t		n_1
Stage 2	Response	Yes (1)	Y_{11}, p_{11}	Y_{10}, p_{10}	Y_r, p_r
		No (0)	Y_{01}, p_{01}	Y_{00}, p_{00}	
	Total		Y_t, p_t		n_2

Type I (or II) error rate may be at an unacceptable level. Bryant and Day (1995) discussed two methods to determine design parameters: The first method deals with a nuisance parameter ϕ, the odds ratio associating the response and toxicity, using a minimax approach; the second assumes independence between the response and toxicity ($\phi = 1$). Conaway and Petroni (1995) presented a method to use ϕ in the design of group sequential trials with bivariate outcomes. For a given ϕ, they chose three quantities, α, γ, and β to control the error rates, where α and γ are for the Type I error and β is for the Type II error. In particular, α, called the "local" bound, is a bound for the Type I error rate at the point (p_{r_0}, p_{t_0}), whereas γ is the maximum Type I error rate over the entire H_0. A disadvantage of using ϕ is that if the true odds ratio value is considerably different from the prespecified one, then the significance level of tests can be adversely affected to an unacceptable level. To reduce the sensitivity of error rates due to misspecifying ϕ, Wu and Liu (2007) proposed an adaptive procedure, which allows the sample size to be re-estimated based on the observed odds ratio. Their procedure is robust against the odds ratio assumption but not clear whether it controls the maximum Type I and II error rates over the entire two hypothesis spaces. Jin (2007) successfully controlled the maximum Type I error rate of a special rejection region (6) through two marginal distributions. However, as shown in Table 2, his attempt to control the minimum power using the nuisance parameter ϕ is not successful. Ray and Rai (2011) developed two-stage designs for hypotheses (1) using the prespecified correlation ρ. But there is no guarantee that their tests are of level-α. Furthermore, all of the above designs are not flexible as they only permit an early termination due to either a noneffective or unsafe treatment in Stage 1. Yu, Kepnerb, and Bundy (2007) proposed a method which includes a relationship value τ for the dependent structure of the two variables. Their method has the same problem as using ϕ. Song (2015) considered two sets of simple hypotheses, one for p_r and the other for p_t, and derived tests using an unreasonable assumption that the number of improved patients is independent of the number of nontoxic patients.

To overcome these drawbacks, we develop a large family of exact tests that strictly control both the maximum Type I and II error rates under a realistic multinomial distribution with considerations of bivariate endpoints and a more general early termination in Stage 1. In Section 2, we first briefly review the two-stage design methodology with bivariate endpoints, then derive level-α tests for (1) by showing that all proposed tests in this paper have nondecreasing power functions in p_r and p_t, respectively. The maximum Type I error rate is achieved when (p_r, p_t, p_{11}) is equal to one of the two points $(p_{r_0}, 1, p_{r_0})$ and $(1, p_{t_0}, p_{t_0})$, displayed as $\underline{p}^*_{0r}$ and $\underline{p}^*_{0t}$ in Figure 1. This result is established by Jin (2007) for a special rejection region (6). We show it is also true for a much more general region (5). The minimum power is achieved at some point on the line $\{(p_r, p_t, p_{11}) = (p_{r_1}, p_{t_1}, p_{11}) : p_{11} \in [\max(0, p_{r_1} + p_{t_1} - 1), \min(p_{r_1}, p_{t_1})]\}$, and this line is displayed as the point $\underline{p}^*_A$ in Figure 1. These three parameter points are on the boundaries of H_0 and H_A in (1) and greatly simplify the computation for the maximum Type I error rate and the minimum power for a given test. In Section 3, we construct optimal designs with exact probability calculation based on multinomial distribution and discuss some examples. In Section 4, an optimization algorithm is discussed. In Section 5, we extend the results of two-stage designs to m-stage designs for $m > 2$. Section 6 contains a discussion. All proofs are given in the Appendix.

2 | EXACT TESTS FOR (1) IN TWO-STAGE DESIGNS

Consider a two-stage design with up to two samples of sizes n_1 (from Stage 1) and n_2 (from Stage 2). All observations are summarized in Table 1, where $X_{ij} = \sum_{v=1}^{n_1} Z_{ijv}$ and $Y_{ij} = \sum_{v=n_1+1}^{n_1+n_2} Z_{ijv}$ are the numbers of observations with response outcome i (0 or 1) and nontoxicity outcome j (0 or 1) in Stage 1 and Stage 2, respectively. Then

$$\underline{X} = (X_{11}, X_{10}, X_{01}) \sim Multinomial(n_1, \underline{p}), \underline{Y} = (Y_{11}, Y_{10}, Y_{01}) \sim Multinomial(n_2, \underline{p}), \tag{3}$$

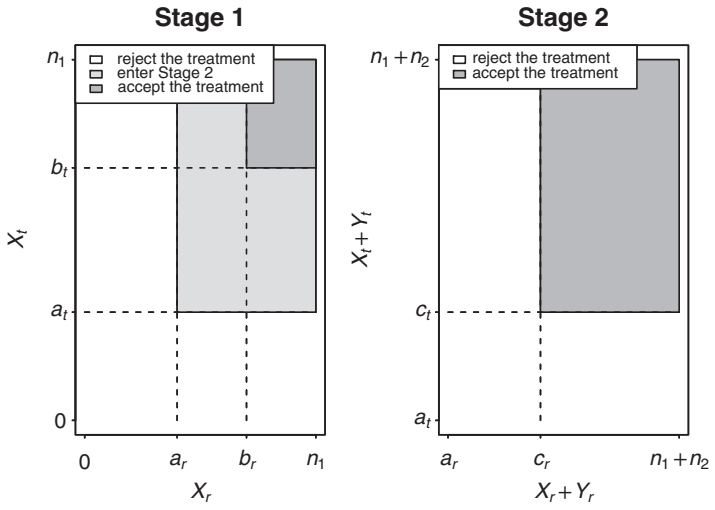

FIGURE 2 A two-stage design with consideration of both response and toxicity. The dark gray areas are the rejection region R_C in (8)

where $\underline{p} = (p_{11}, p_{10}, p_{01})$. Also let $X_r = X_{11} + X_{10}$ and $X_t = X_{11} + X_{01}$ be the numbers of response and nontoxicity in Stage 1, respectively. Similarly, let $Y_r = Y_{11} + Y_{10}$ and $Y_t = Y_{11} + Y_{01}$ in Stage 2.

Consider the following general case for two-stage designs. An early termination of the trial (i.e., $n_2 = 0$) occurs if (i) the treatment is noneffective or unsafe (i.e., the termination is due to futility or unsafety) or (ii) the treatment is effective and safe (i.e., the termination is due to superiority and safety). If an early termination does not occur, then we enter Stage 2 and observe $\underline{Y}$, so, $n_2 > 0$. The design is specified by a vector of 12 integer-valued parameters

$$Q = (n_1, a_r, a_t, b_r, b_t, n_2, c_{r_1}, c_{t_1}, c_{r_2}, c_{t_2}, c_{r_3}, c_{t_3}) \tag{4}$$

introduced below. The trial proceeds as follows:

- Stage 1. Select the first random sample of n_1 patients and record X_r and X_t.
  - If $X_r \geq b_r$ and $X_t \geq b_t$, accept the treatment and terminate the trial due to superiority and safety.
  - The trial enters Stage 2 on three mutually exclusive events in Stage I:
    * if $a_r \leq X_r < b_r$ and $X_t \geq b_t$ (denote this event by $\mathcal{S}_1$);
    * if $X_r \geq b_r$ and $a_t \leq X_t < b_t$ (denote this event by $\mathcal{S}_2$);
    * if $a_r \leq X_r < b_r$ and $a_t \leq X_t < b_t$ (denote this event by $\mathcal{S}_3$).
  - Otherwise, the trial is terminated in Stage 1 due to futility or unsafety.
- Stage 2. When the trial enters Stage 2, select the second sample of n_2 patients and record Y_r and Y_t.
  - Accept the treatment due to superiority and safety on each of three mutually exclusive events if $X_r + Y_r \geq c_{r_i}$ and $X_t + Y_t \geq c_{t_i}$ for $i = 1, 2, 3$.
  - Otherwise, reject the treatment due to futility or unsafety.

Therefore, the rejection region for (1) is

$$R = \bigcup_{i=1}^{3} \{\mathcal{S}_i \cap \{X_r + Y_r \geq c_{r_i}, X_t + Y_t \geq c_{t_i}\}\} \cup \{X_r \geq b_r, X_t \geq b_t\}, \tag{5}$$

where $a_r, a_t, b_r, b_t, c_{r_i}, c_{t_i}$ for $i = 1, 2, 3$ are 10 constants to determine a two-stage design, as well as n_1 and n_2, for given nominal levels α and β. Naturally, we require: $a_r < b_r \leq c_{r_i} \leq c_{r_3}$ and $a_t < b_t \leq c_{t_i} \leq c_{t_3}$ for $i = 1, 2$. Figure 2 contains R when c_{r_i}s are equal to a constant c_r for $i = 1, 2, 3$ and $c_{t_i} = c_t$ similarly. Later in this section, we find out where the probability $P(R)$ achieves its maximum in H_0 (the maximum Type I error rate) and its minimum in H_A (the minimum power and, equivalently, the maximum Type II error rate). In practice, it may not be appropriate to determine all 12 components in Q due to computational complexity. Here are three practical choices:

Case A. Simon (1989) derived optimal two-stage designs, which only allow an early termination due to futility, for a single variable of the response. Here, a parallel result in the case of bivariate variables is to choose $b_t = b_r = n_1 + 1$ and write $c_{r_3} = c_r, c_{t_3} = c_t$. This only allows an early termination due to futility or unsafety in Stage 1. Then R in (5) reduces to

$$R_A = \{X_r \geq a_r, X_t \geq a_t, X_r + Y_r \geq c_r, X_t + Y_t \geq c_t\}, \tag{6}$$

which involves four constants, a_r, a_t, c_r, and c_t. Similar results were discussed in Bryant and Day (1995), Conaway and Petroni (1995), and Jin (2007).

Case B. If an early termination due to superiority and safety is also of interest (such a choice would reduce the expected total sample size of a two-stage design), one may choose $c_{t_1} = b_t$ and $c_{r_2} = b_r$. The reason for choosing $c_{t_1} = b_t$, for example, is that on S_1 a large X_t is observed in Stage 1 (indicating a safe treatment), then there is no need to observe Y_t anymore in Stage 2. Further, let $c_{r_1} = c_{r_3} = c_r$ and $c_{t_2} = c_{t_3} = c_t$. Then R in (5) reduces to

$$\begin{aligned} R_B = {} & \{X_r \geq b_r, X_t \geq b_t\} \cup \{S_1 \cap \{X_r + Y_r \geq c_r\}\} \\ & \cup \{S_2 \cap \{X_t + Y_t \geq c_t\}\} \cup \{S_3 \cap \{X_r + Y_r \geq c_r, X_t + Y_t \geq c_t\}\}, \end{aligned} \tag{7}$$

which involves six constants, a_r, a_t, b_r, b_t, c_r, and c_t. If $b_r = b_t = n_1 + 1$, then R_B reduces to R_A. In Section 3, we focus on the construction of R_B.

Case C. On S_1, for example, Case B does not observe Y_t in Stage 2 due to a large X_t in Stage 1. However, as far as the data are effectively collected from each patient, it may still be desirable to observe Y_t since we are in Stage 2 anyway. Let $c_{r_i} = c_r$ and $c_{t_i} = c_t$ for $i = 1, 2, 3$. Then R in (5) reduces to

$$R_C = \bigcup_{i=1}^{3} \{S_i \cap \{X_r + Y_r \geq c_r, X_t + Y_t \geq c_t\}\} \cup \{X_r \geq b_r, X_t \geq b_t\}, \tag{8}$$

which involves six constants, a_r, a_t, b_r, b_t, c_r, and c_t. This region is given in Figure 2. If $c_t = b_t$ and $c_r = b_r$, then R_C reduces to R_B.

In summary, the test R in (5) includes the tests R_A, R_B, and R_C; R_B includes R_A; and R_C includes R_B. Both R_B and R_C allow an early stopping due to superiority and safety, so they are more flexible than R_A. Between R_B and R_C, R_B requires less data as explained above; while R_C may draw more reliable inferences since it utilizes more data. Lastly, R is more flexible than R_C by using different constants c_{r_i}s and c_{t_i}s, which depend on the outcomes S_i in Stage 1. We next state three results on the monotonicity of probability function $P(R)$, the maximum Type I error rate, and the minimum power of the test R.

Theorem 2.1. *Consider the rejection region R in* (5) *with $c_{r_i} \leq c_{r_3}$ and $c_{t_i} \leq c_{t_3}$ for $i = 1, 2$. Let $\underline{p}^* = (p_r, p_t, p_{11})$, which is different from $\underline{p} = (p_{11}, p_{10}, p_{01})$ introduced after* (3). *Define $f(\underline{p}^*) = P(R)$ as a function of p_r, p_t, and p_{11}. Then,*

(i) for any fixed p_t and p_{11}, $f(\underline{p}^*)$ is a nondecreasing function for p_r;

(ii) for any fixed p_r and p_{11}, $f(\underline{p}^*)$ is a nondecreasing function for p_t.

Theorem 2.2. *The maximum Type I error rate of the test R, $\max_{H_0} P(R)$, is achieved at either $\underline{p}^*_{0r} = (p_{r_0}, 1, p_{r_0})$ or $\underline{p}^*_{0t} = (1, p_{t_0}, p_{t_0})$. That is,*

$$\max_{H_0} P(R) = \max_{\underline{p}^* \in H_0} f(\underline{p}^*) = \max\{f(\underline{p}^*_{0r}), f(\underline{p}^*_{0t})\}.$$

Theorem 2.3. *The minimum power of the test R, $\min_{H_A} P(R)$, is achieved at some point on the line $\underline{p}^*_A = (p_{r_1}, p_{t_1}, p_{11})$ for $p_{11} \in I_A = [\max(0, p_{r_1} + p_{t_1} - 1), \min(p_{r_1}, p_{t_1})]$. That is,*

$$\min_{H_A} P(R) = \min_{\underline{p}^* \in H_A} f(\underline{p}^*) = \min_{p_{11} \in I_A} f(\underline{p}^*_A).$$

Theorem 2.1 is needed to establish Theorem 2.3. However, different from the univariate case, the former does not directly imply the latter. Following Theorem 2.2, any design parameter vector Q in (4) is evaluated by checking whether the Type I error rates for the associated test R at $\underline{p}^*_{0r}$ and $\underline{p}^*_{0t}$ are no larger than the given level α. Therefore, finding the maximum on a three-dimensional set H_0 is simplified to finding the maximum on only two points. Theorem 2.3 plays a similar role on power. However, the search for the minimum power on a three-dimensional set H_A is replaced by a search on the one-dimensional set

I_A, which is denoted by $\underline{p}_A^*$ in Figure 1. If the test R with the given vector Q has a maximum Type I error rate no larger than α and a minimum power no smaller than $1-\beta$, then this Q is kept; otherwise, it is discarded. The two theorems ensure the test R having the correct Type I and II error rates and result in reliable statistical inferences. Finally, since we keep finitely many Qs, the Q with the smallest expected total sample size exists and can be found through numerical search.

3 | OPTIMAL TWO-STAGE DESIGNS

The previous section discusses the mathematical foundation to assess the test R in (5). We now provide guidance on how to select optimal two-stage designs for practice by providing a variation of Q in (4). Here, an early termination due to superiority and safety is allowed; and it is reasonable to assume that Y_r (or Y_t) in Stage 2 is not needed if a large value of X_r (or X_t) is observed in Stage 1. Then, we focus on R_B in (7) and need to determine a set of simplified design parameters

$$Q_B = (n_1, a_r, a_t, b_r, b_t, n_2, c_r, c_t). \tag{9}$$

The goal of this section is to find an optimal design, denoted by Q_B^{opt}, so that the maximum Type I and II error rates are controlled at given α and β, respectively, the total sample size is no larger than a given upper limit N^*, and the expected total sample size is minimized.

Let PET be the probability of early termination in Stage 1. It is a function of $\underline{p}^*$ and is equal to

$$PET(\underline{p}^*) = P(\{X_r < a_r\} \cup \{X_t < a_t\}) + P(X_r \geq b_r, X_t \geq b_t).$$

Let N be the (random) total number of patients in a two-stage design. We require N no larger than N^*. The expected total sample size EN, also a function of $\underline{p}^*$, is

$$EN(\underline{p}^*) = n_1 + (1 - PET(\underline{p}^*))n_2.$$

Simon (1989) suggested to minimize EN at $p_r = p_{r_0}$, the boundary of null hypothesis in (2), with control of the Type I and II error rates in his univariate case. If following his suggestion, find the Q_B^{opt} that controls the maximum Type I and II error rates and minimizes

$$EN_0 \overset{def}{=} \max\{EN(\underline{p}_{0r}^*), EN(\underline{p}_{0t}^*)\}, \tag{10}$$

where $\underline{p}_{0r}^*$ and $\underline{p}_{0t}^*$ are introduced in Theorem 2.2. More precisely, let $\mathcal{B}$ be the set of tests R_B with the design parameter vector Q_B and the total sample size $N \leq N^*$ such that

$$\begin{aligned} &\max_{H_0} P(R_B) = \max P(R_B \mid \underline{p}^* = \underline{p}_{0r}^* \text{ or } \underline{p}_{0t}^*) \leq \alpha \\ &\min_{H_A} P(R_B) = \min_{p_{11} \in I_A} P(R_B \mid \underline{p}^* = \underline{p}_A^*) \geq 1-\beta. \end{aligned} \tag{11}$$

We need to determine a test R_B^{opt} in the set of tests $\mathcal{B}$ with the design parameter vector $Q_B^{opt} = (n_1^{opt}, a_r^{opt}, a_t^{opt}, b_r^{opt}, b_t^{opt}, n_2^{opt}, c_r^{opt}, c_t^{opt})$ and the total sample size N^{opt} such that

$$EN_0^{opt} = \min_{\mathcal{B}}\{EN_0\}. \tag{12}$$

The following three facts greatly simplify the calculation of $\max_{H_0} P(R_B)$ and $\min_{H_A} P(R_B)$ in (11):

(i) $P(R_B \mid \underline{p}^* = \underline{p}_{0r}^*) = P(X_{11} \geq b_r) + P(a_r \leq X_{11} < b_r, X_{11} + Y_{11} \geq c_r)$ for $X_{11} \sim Binomial(n_1, p_{r_0})$ and $Y_{11} \sim Binomial(n_2, p_{r_0})$ as $X_{10} = X_{00} = Y_{10} = Y_{00} \equiv 0$ when $\underline{p}^* = \underline{p}_{0r}^*$.

(ii) $P(R_B \mid \underline{p}^* = \underline{p}_{0t}^*) = P(X_{11} \geq b_t) + P(a_t \leq X_{11} < b_t, X_{11} + Y_{11} \geq c_t)$ for $X_{11} \sim Binomial(n_1, p_{t_0})$ and $Y_{11} \sim Binomial(n_2, p_{t_0})$ as $X_{01} = X_{00} = Y_{01} = Y_{00} \equiv 0$ when $\underline{p}^* = \underline{p}_{0t}^*$.

TABLE 2 The minimum power of Case A in Table 2, Jin (2007), when $(p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1}) = (0.5, 0.7, 0.7, 0.9)$ and $1 - \beta = 0.8$

Method	n_1	a_r	a_t	n_2	c_r	c_t	Minimum power
Conaway and Petroni	11	6	9	26	22	29	0.794 (<0.8)
Jin	15	9	12	29	25	36	0.793 (<0.8)
	20	12	16	16	21	30	0.796 (<0.8)

TABLE 3 The maximum Type I error rate and minimum power of Case C in Table 2, Yu et al. (2007), when $(\alpha, 1 - \beta) = (0.05, 0.8)$

Method	$(p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1})$	n_1	a_r	a_t	b_r	b_t	n_2	c_r	c_t	Maximum Type I error rate	Minimum power
I	(0.3, 0.60, 0.5, 0.80)	14	0	0	8	11	14	12	20	0.2029 (>0.05)	0.747 (<0.8)
	(0.1, 0.65, 0.3, 0.85)	10	0	0	4	8	9	4	15	0.3022 (>0.05)	0.741 (<0.8)
II	(0.3, 0.60, 0.5, 0.80)	13	4	8	14	14	13	9	19	0.3268 (>0.05)	0.796 (<0.8)
	(0.4, 0.65, 0.6, 0.85)	13	5	8	14	14	12	11	20	0.3780 (>0.05)	0.786 (<0.8)
	(0.4, 0.65, 0.6, 0.85)	14	6	9	15	15	13	14	21	0.1367 (>0.05)	0.742 (<0.8)

(iii) Numerical calculations show that, in general, the minimum power, $\min_{H_A} P(R_B)$, is not always achieved at a fixed parameter point. However, in many cases including those in Tables 4 and 5 later, the minimum power is achieved at a fixed point $\underline{p}^*_{A0} = (p_{r_1}, p_{t_1}, \max(0, p_{r_1} + p_{t_1} - 1))$, the starting point of line $\underline{p}^*_A$ given in Theorem 2.3. That is,

$$P(R_B \mid \underline{p}^* = \underline{p}^*_{A0}) = \min_{p_{11} \in I_A} P(R_B \mid \underline{p}^* = \underline{p}^*_A). \tag{13}$$

Benefit from (13) in practice, we first compute the minimum power using the left-hand side of (13) to find an optimal design Q_B^{opt}. Second, we check numerically whether (13) is true for this Q_B^{opt}. If yes, then the Q_B^{opt} is truly optimal (see Lemma 6.4 in the Appendix). If no, then the Q_B^{opt} may not be optimal, and we have to compute the minimum power using the right-hand side of (13) to find an optimal design.

Conaway and Petroni (1995) and Jin (2007) controlled the power of Case A at least 0.8 by specifying a value of the odds ratio ϕ. However, the power should be minimized over H_A including all ϕs, which, in particular, should include the specified value and others. In consequence, their minimum power may be smaller than the desired level 0.8 as shown in Table 2. Similarly, Yu et al. (2007) computed the Type I error rate and power of Case C by specifying a value of the relationship parameter $\tau = (p_r p_t - p_{11})/(p_{11} - p_r(1 - p_t + p_r))$; see Equation (1) in Yu et al. (2007). We find in Table 3 that their tests do not necessarily control the maximum Type I error rate over H_0 at α and the minimum power over H_A at $1 - \beta$. In contrast, we next construct optimal two-stage designs satisfying the two constraints in (11).

Example 1.1 (continued). A numerical search yields an optimal two-stage design $Q_B^{opt} = (n_1^{opt}, a_r^{opt}, a_t^{opt}, b_r^{opt}, b_t^{opt}, n_2^{opt}, c_r^{opt}, c_t^{opt}) = (29, 14, 14, 18, 18, 34, 32, 32)$. The corresponding test R_B^{opt} has a maximum Type I error rate 0.04892 and a minimum power 0.8034. Also, it has an expected total sample size $EN_0^{opt} = 36.5013$, much smaller than $n_1^{opt} + n_2^{opt} = 63$ and $N^* = 65$. We also provide the Type I error rate and power for this case in Figure 3. The original Type I error rate and power are functions of three variables, p_r, p_t, and p_{11}. To make a three-dimensional plot, the number of independent variables must reduce to two. Similar to I_A in Theorem 2.3, let $I = [\max(0, p_r + p_t - 1), \min(p_r, p_t)]$. Then, the Type I error rate in Figure 3 is defined to be

$$\max_{p_{11} \in I} f(p_r, p_t, p_{11})$$

for each (p_r, p_t), which is an upper bound for $f(p_r, p_t, p_{11})$. The power function in Figure 3 is equal to

$$\min_{p_{11} \in I} f(p_r, p_t, p_{11})$$

for each (p_r, p_t), which is a lower bound for $f(p_r, p_t, p_{11})$. The Type I error rate does not go above 0.05 $(= \alpha)$, and the power function does not go below 0.8 $(= 1 - \beta)$, both confirming Theorems 2.2 and 2.3.

Table 4 contains more optimal two-stage designs for a variety of hypothesized values satisfying $(p_{r_0}, p_{r_1}) = (p_{t_0}, p_{t_1})$, two sets of error probabilities $(\alpha, \beta) = (0.05, 0.2)$ or $(0.05, 0.1)$, and some predetermined N^*. The key point is that the expected total sample size EN_0^{opt} is much smaller than $n_1 + n_2$.

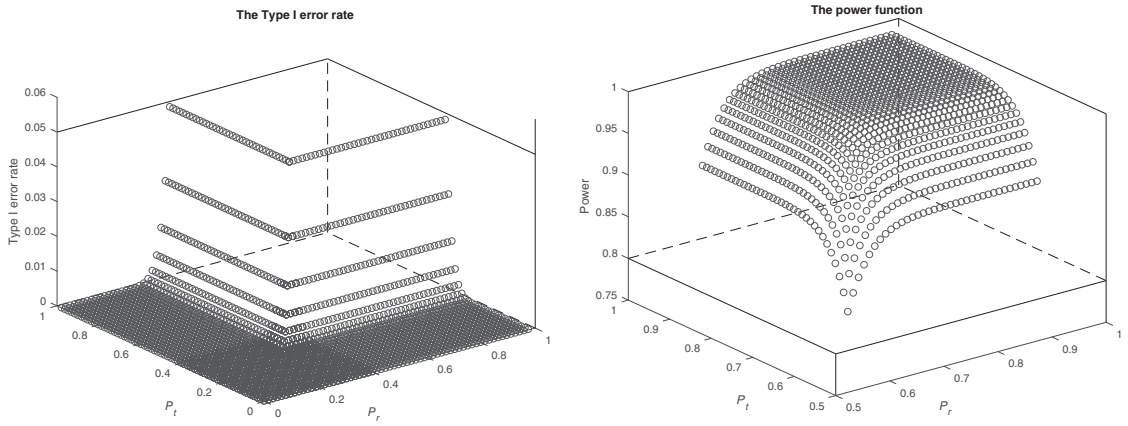

FIGURE 3 The Type I error rate of R_B^{opt} in H_0 (left) and the power function of R_B^{opt} in H_A (right) for $(\alpha, \beta, p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1}) = (0.05, 0.2, 0.4, 0.4, 0.6, 0.6)$ and $Q_B^{opt} = (29, 14, 14, 18, 18, 34, 32, 32)$

TABLE 4 Optimal two-stage designs Q_B^{opt} in Case B when EN_0 is minimized for $(\alpha, \beta) = (0.05, 0.2)$ or $(0.05, 0.1)$ and for given N^*

$(p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1})$	n_1^{opt} b_t^{opt}	a_r^{opt} n_2^{opt}	a_t^{opt} c_r^{opt}	b_r^{opt} c_t^{opt}	EN_0^{opt} N^*	$\max_{H_0} P(R_B)$	$\min_{H_A} P(R_B)$	β
(0.2, 0.2, 0.4, 0.4)	22	6	6	10	30.3591	0.04899	0.8098	0.2
	10	32	16	16	65			
	28	7	7	13	38.8835	0.04983	0.9013	0.1
	13	34	18	18	70			
(0.3, 0.3, 0.5, 0.5)	24	9	9	15	34.6852	0.04992	0.8067	0.2
	15	39	25	25	65			
	30	10	10	16	46.1928	0.04638	0.9026	0.1
	16	40	28	28	70			
(0.4, 0.4, 0.6, 0.6)	29	14	14	18	36.5013	0.04892	0.8034	0.2
	18	34	32	32	65			
	35	15	15	21	49.0756	0.04984	0.9003	0.1
	21	34	35	35	70			
(0.5, 0.5, 0.7, 0.7)	21	12	12	17	33.4721	0.04998	0.8020	0.2
	17	38	36	36	65			
	36	20	20	25	45.7169	0.04826	0.9021	0.1
	25	33	42	42	70			
(0.6, 0.6, 0.8, 0.8)	18	12	12	17	28.8157	0.04933	0.8016	0.2
	17	29	34	34	65			
	26	17	17	22	37.7993	0.04900	0.9014	0.1
	22	33	42	42	70			

For fixed n_1 and n_2, the following result on monotonicity can eliminate many cases of $(a_r, a_t, b_r, b_t, c_r, c_t)$ from the search for Q_B^{opt}.

Theorem 3.1. *For a fixed parameter point (p_r, p_t, p_{11}), two fixed sample sizes n_1 and n_2, and R_B in (7), consider $g_{R_B}(a_r, a_t, b_r, b_t, c_r, c_t) = P(R_B)$ as a function of $(a_r, a_t, b_r, b_t, c_r, c_t)$. Also consider $g_{EN_0}(a_r, a_t, b_r, b_t) = EN_0$ as a function of (a_r, a_t, b_r, b_t). Then,*

(i) $g_{R_B}(a_r, a_t, b_r, b_t, c_r, c_t)$ is nonincreasing in each of its arguments when the others are fixed;

(ii) $g_{EN_0}(a_r, a_t, b_r, b_t)$ is nonincreasing in each of a_r and a_t and is nondecreasing in each of b_r and b_t when the other arguments are fixed.

TABLE 5 Optimal two-stage designs Q_B^{opt} in Case B when EN_A is minimized for $(\alpha, \beta) = (0.05, 0.2)$ or $(0.05, 0.1)$ and for given N^*

$(p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1})$	n_1^{opt} b_t^{opt}	a_r^{opt} n_2^{opt}	a_t^{opt} c_r^{opt}	b_r^{opt} c_t^{opt}	EN_A^{opt} N^*	$\max_{H_0} P(R_B)$	$\min_{H_A} P(R_B)$	β
(0.3, 0.3, 0.5, 0.5)	35	12	12	16	44.6306	0.04905	0.8026	0.2
	16	21	24	24	65			
	39	11	11	18	53.0135	0.04981	0.9001	0.1
	18	27	27	27	70			
(0.3, 0.4, 0.5, 0.6)	37	13	18	17	46.1084	0.04817	0.8040	0.2
	21	20	24	30	65			
	39	13	17	19	58.0102	0.04932	0.9048	0.1
	22	31	28	36	70			
(0.4, 0.4, 0.6, 0.6)	37	17	17	21	46.9503	0.04886	0.8057	0.2
	21	20	30	30	65			
	39	17	17	23	59.8520	0.04898	0.9020	0.1
	23	30	35	35	70			
(0.4, 0.5, 0.6, 0.7)	33	15	19	19	46.6764	0.04976	0.8013	0.2
	22	28	32	39	65			
	39	16	21	22	54.3337	0.04996	0.9018	0.1
	26	31	36	43	70			
(0.5, 0.6, 0.7, 0.8)	31	17	19	21	42.2082	0.04898	0.8002	0.2
	24	22	34	39	65			
	38	21	23	25	49.6404	0.04871	0.9006	0.1
	29	29	42	48	70			

To illustrate how to apply this result, suppose $(\alpha, \beta, p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1}, n_1, n_2) = (0.05, 0.2, 0.5, 0.5, 0.7, 0.7, 22, 39)$. Consider $Q_B = (22, 13, 13, 19, 19, 39, 36, 36)$, where Q_B is defined in (9). We compute the minimum power $g_{R_B}(a_r, a_t, b_r, b_t, c_r, c_t) = g_{R_B}(13, 13, 19, 19, 36, 36) = 0.796 < 1 - \beta$. Then, any $Q'_B = (22, a_r, a_t, b_r, b_t, 39, c_r, c_t)$ with $a_r \geq 13, a_t \geq 13, b_r \geq 19, b_t \geq 19, c_r \geq 36, c_t \geq 36$ is eliminated in the search for Q_B^{opt} since such a Q'_B has a minimum power less than that of Q_B as claimed in part (i) of Theorem 3.1. Also, consider another $Q_B = (22, 12, 12, 17, 17, 39, 38, 38)$ with a maximum type I error rate 0.038, a minimum power 0.812, and $g_{EN_0}(a_r, a_t, b_r, b_t) = g_{EN_0}(12, 12, 17, 17) = 37.891$. Then, any $Q'_B = (22, a_r, a_t, b_r, b_t, 39, c_r, c_t)$ with $a_r \leq 12, a_t \leq 12, b_r \geq 17, b_t \geq 17$ is eliminated in the search for Q_B^{opt} since such a Q'_B has a larger EN_0 than Q_B as claimed in part (ii) of Theorem 3.1.

Pointed out by a referee, the expected total sample size on H_A may also be of interest since we allow an early stopping for superiority and safety. Suppose the minimum power on H_A is achieved at a point, say $\underline{p}_{Am}^*$. As mentioned before, $\underline{p}_{Am}^* = \underline{p}_{A0}^*$ in many cases. We wish to find an optimal design Q_B^{opt} that satisfies (11) and $n_1 + n_2 \leq N^*$ and minimizes the expected total sample size at $\underline{p}_{Am}^*$, denoted by EN_A. Table 5 contains such designs. Tables 4 and 5 show that optimal designs depend on which of the EN_0 and EN_A is minimized.

4 | AN ALGORITHM TO CONSTRUCT Q_B^{opt}

In this section, we provide the algorithm details to derive the optimal design Q_B^{opt} that minimizes EN_0 for given N^*, an upper limit of $n_1 + n_2$. Other optimal designs can be computed similarly.

The determination of Q_B^{opt} is done by numerical search. Hypothetically speaking, (a) we, for a given Q_B, compute the Type I error rate at $\underline{p}_{0r}^*$ and $\underline{p}_{0t}^*$, the minimum power at $\underline{p}_A^*$ and EN_0; (b) run eight loops for each component of Q_B in a range and keep those Q_Bs satisfying (11) and $n_1 + n_2 \leq N^*$; (c) sort these Q_Bs by EN_0. Then the one with the smallest EN_0 is Q_B^{opt}. The optimal design Q_B^{opt} with the design parameter vector $(n_1^{opt}, a_r^{opt}, a_t^{opt}, b_r^{opt}, b_t^{opt}, n_2^{opt}, c_r^{opt}, c_t^{opt})$ can be computed by following Steps (I)–(IV):

Step I. Set the values, $(N^*, \alpha, \beta, p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1})$, where $N^*(\geq n_1 + n_2)$ is a predetermined upper limit for the total number of patients, $(p_{r_0}, p_{t_0}, p_{r_1}, p_{t_1})$ are the hypothesized values in H_0 and H_A, and α and β are the levels of the Type I and II error rates, respectively.

Step II. Run over all situations for Q_B. That is, for each $Q_B = (n_1, a_r, a_t, b_r, b_t, n_2, c_r, c_t)$, where

$$n_1 \in [1, N^*], \quad n_2 \in [0, N^* - n_1], \quad a_r \in [0, n_1], a_t \in [0, n_1],$$

$$b_r \in [a_r + 1, n_1 + 1], \quad b_t \in [a_t + 1, n_1 + 1], \quad c_r \in [b_r, n_1 + n_2], \quad c_t \in [b_t, n_1 + n_2]$$

(eight loops), calculate its maximum Type I error rate and minimum power:

$$\max\{P(R_B|\underline{p}^* = \underline{p}^*_{0r}), P(R_B|\underline{p}^* = \underline{p}^*_{0t})\} \text{ and } \min P(R_B|\underline{p}^* = \underline{p}^*_A).$$

The computation of the above quantities can be greatly simplified by using the three facts after (12). If a strong computing power is available, jump to Step III; otherwise, follow the steps below to filter out many disqualified Q_Bs for (11). That is, reduce the range of the loops listed above.

Step II-1 (Smaller upper limits for a_r and a_t). The range of a_r is reduced from $[0, n_1]$ to $[0, U_{a_r}]$, where

$$U_{a_r} = \max\{a_r : P(X_r \geq a_r|\underline{p}^* = \underline{p}^*_{Am}) \geq 1 - \beta\}$$

and $\underline{p}^*_{Am}$ is defined before Table 5. Any a'_r larger than U_{a_r} yields a disqualified Q_B because $P(R_B|\underline{p}^* = \underline{p}^*_{Am}) \leq P(X_r \geq a'_r|\underline{p}^* = \underline{p}^*_{Am}) < 1 - \beta$. Similarly, the reduced range of a_t is $[0, U_{a_t}]$ for

$$U_{a_t} = \max\{a_t : P(X_t \geq a_t|\underline{p}^* = \underline{p}^*_{Am}) \geq 1 - \beta\}.$$

Step II-2 (Larger lower limits for b_r and b_t). The reduced range of b_r is $[\max\{a_r + 1, L_{b_r}\}, n_1 + 1]$ for

$$L_{b_r} = \min\{b_r : P(X_r \geq b_r|\underline{p}^* = \underline{p}^*_{0r}) \leq \alpha\}.$$

Any b'_r smaller than L_{b_r} yields a disqualified Q_B because $P(R_B|\underline{p}^* = \underline{p}^*_{0r}) = P(X_r \geq b'_r|\underline{p}^* = \underline{p}^*_{0r}) + P(a_r \leq X_r < b'_r, X_r + Y_r \geq c_r|\underline{p}^* = \underline{p}^*_{0r}) > P(X_r \geq b'_r|\underline{p}^* = \underline{p}^*_{0r}) > \alpha$. Similarly, the reduced range of b_t is $[\max\{a_t + 1, L_{b_t}\}, n_1 + 1]$, where

$$L_{b_t} = \min\{b_t : P(X_t \geq b_t|\underline{p}^* = \underline{p}^*_{0t}) \leq \alpha\}.$$

Step II-3 (Smaller upper limits for c_r and c_t). The reduced range of c_r is $[b_r, U_{c_r}]$, where

$$U_{c_r} = \max\{c_r : P(X_r \geq b_r|\underline{p}^* = \underline{p}^*_{Am}) + P(a_r \leq X_r < b_r, X_r + Y_r \geq c_r|\underline{p}^* = \underline{p}^*_{Am}) \geq 1 - \beta\}$$

for fixed a_r and b_r. The reason of choosing this U_{c_r} is similar to that for U_{a_r}. Also, $c_t \in [b_t, U_{c_t}]$, where

$$U_{c_t} = \max\{c_t : P(X_t \geq b_t|\underline{p}^* = \underline{p}^*_{Am}) + P(a_t \leq X_t < b_t, X_t + Y_t \geq c_t|\underline{p}^* = \underline{p}^*_{Am}) \geq 1 - \beta\}.$$

Step II-4 (more reductions). Our experiences and Simon (1989) suggest that (i) $n_1 \in [\frac{N_l}{2} - m_1, \frac{N^*}{2} + m_2]$, where $N_l = \underline{p}(1 - \underline{p})[\frac{Z_{1-\alpha} + Z_{1-\beta}}{(p_1 - p_0)}]^2$ and for $p_0 = \frac{1}{2}(p_{r_0} + p_{t_0})$, $p_1 = \frac{1}{2}(p_{r_1} + p_{t_1})$, $\underline{p} = \frac{1}{2}(p_0 + p_1)$ and some $m_1 \in [0, 5]$ and $m_2 \in [0, 5]$; (ii) $a_r \in [n_1 p_{r_0}, U_{a_r}]$, $a_t \in [n_1 p_{t_0}, U_{a_t}]$; (iii) $c_r \in [\max\{(n_1 + n_2)p_{r_0}, b_r\}, U_{c_r}]$, $c_t \in [\max\{(n_1 + n_2)p_{t_0}, b_t\}, U_{c_t}]$; (iv) $n_2 \in [\min\{n_1, N^* - n_1\} - m_3, \min\{\frac{3}{4}N^*, N^* - n_1\}]$ for some $m_3 \in [0, 10]$.

Step II-5. If $(p_{r_0}, p_{r_1}) = (p_{t_0}, p_{t_1})$, it seems reasonable to assume $a_r = a_t$, $b_r = b_t$ and $c_r = c_t$. Then, the number of loops in computation drops to five from eight. Interestingly, the final Q_B^{opt} always has $a_r^{opt} = a_t^{opt}$, $b_r^{opt} = b_t^{opt}$ and $c_r^{opt} = c_t^{opt}$, as shown in Tables 4 and 5.

Step III. Record those Q_Bs satisfying (11) in **Step II** and compute EN_0 using (10) for each Q_B.

Step IV. Sort all Q_Bs in **Step III** by EN_0. Then the Q_B with the smallest EN_0 is equal to Q_B^{opt}.

One should be aware of that Steps II-4 and II-5 can reduce the computation time, but unlike Steps II-1 through II-3, the two are not established by mathematical proofs. The resultant designs following the two steps are not guaranteed to be always optimal. However, our extensive calculations show that they are optimal or close to the optimal ones, including those in Tables 4 and 5. If we enlarge the range of loops in Step II-4, for example, choose larger m_is, then the resultant designs should be optimal. In

the computation of Tables 4 and 5, $m_1 = m_2 = 5$ and $m_3 = 10$ are chosen. If only the optimal designs are needed and a strong computing power is available, one can find them using Step II without Steps II-4 and II-5.

5 | OPTIMAL MULTISTAGE DESIGNS

In this section, we extend the results to m-stage designs for $m > 2$. When the trial enters Stage k for $k = 1, \ldots, m$, observe $\underline{X}^{(k)} = (X_{11}^{(k)}, X_{10}^{(k)}, X_{01}^{(k)})$, a sufficient statistic for $\underline{p}$ in Stage k, that contains the status of response and nontoxicity in the kth sample of n_k patients. For example, $X_{11}^{(k)}$ is the number of improved and nontoxic patients in Stage k. In particular, $\underline{X}$ and $\underline{Y}$ in (3) are equal to $\underline{X}^{(k)}$ for $k = 1, 2$, respectively. Then,

$$\underline{X}^{(k)} \sim Multinomial(n_k, \underline{p}) \text{ for } \underline{p} = (p_{11}, p_{10}, p_{01}).$$

Let $X_r^{(k)} = X_{11}^{(k)} + X_{10}^{(k)}$ be the number of patients who show improvement, and let $X_t^{(k)} = X_{11}^{(k)} + X_{01}^{(k)}$ be the number of patients who are nontoxic in Stage k. Also let $S_r^{(k)} = \sum_{j=1}^{k} X_r^{(j)}$ be the total number of patients who show improvement and let $S_t^{(k)} = \sum_{j=1}^{k} X_t^{(j)}$ be the total number of patients who are nontoxic up to Stage k. The goal of this section is to derive exact tests for (1) based on $\{S_i^{(k)}\}_{k=1}^{m_0}$ for $i = r, t$ and some $m_0 \le m$, where m_0 is the random number of stages at which the trial ends.

Intuitively, we reject H_0 in (1) if both $S_r^{(k)}$ and $S_t^{(k)}$ are large and fail to reject H_0 if $S_r^{(k)}$ or $S_t^{(k)}$ is small, and in any of these two cases we terminate the trial in Stage k; otherwise (i.e., if one of $S_r^{(k)}$ and $S_t^{(k)}$ is medium and the other one is not small), we enter Stage $k + 1$ and collect more data for a decision.

More precisely, consider two given sequences of integer vectors $(a_r^{(k)}, a_t^{(k)})$ and $(b_r^{(k)}, b_t^{(k)})$ with each of the components monotonically nondecreasing in k, $0 \le a_r^{(k)} < b_r^{(k)} \le N_k + 1, 0 \le a_t^{(k)} < b_t^{(k)} \le N_k + 1$ for $k = 1, \ldots, m-1$ and $0 \le a_r^{(m)} = b_r^{(m)} \le N_m, 0 \le a_t^{(m)} = b_t^{(m)} \le N_m$, where $N_k = \sum_{j=1}^{k} n_j$ for $k = 1, \ldots, m$. An m-stage design, which is a generalization of R_C in (8), for testing (1) is given below. The generalization of R_B in (7) can also be given (see the Supporting Information).

I. In Stage 1, we claim the treatment noneffective or unsafe (i.e., H_0) if $S_r^{(1)} < a_r^{(1)}$ or $S_t^{(1)} < a_t^{(1)}$; claim the treatment effective and safe (i.e., H_A) if $S_r^{(1)} \ge b_r^{(1)}$ and $S_t^{(1)} \ge b_t^{(1)}$, and the trial ends in Stage 1; otherwise, the trial enters Stage 2.

II. Assume the trial enters Stage k for some $k > 1$. We claim the treatment noneffective or unsafe (i.e., H_0) if $S_r^{(k)} < a_r^{(k)}$ or $S_t^{(k)} < a_t^{(k)}$; claim the treatment effective and safe (i.e., H_A) if $S_r^{(k)} \ge b_r^{(k)}$ and $S_t^{(k)} \ge b_t^{(k)}$, and the trial ends in Stage k; otherwise, the trial enters Stage $k + 1$.

III. If the trial enters Stage m, then we claim the treatment noneffective or unsafe (i.e., H_0) if $S_r^{(m)} < b_r^{(m)}$ or $S_t^{(m)} < b_t^{(m)}$; claim the treatment effective and safe (i.e., H_A) if $S_r^{(m)} \ge b_r^{(m)}$ and $S_t^{(m)} \ge b_t^{(m)}$.

The rejection region R_m for (1) is

$$R_m = B_1 \cup \left\{ \cup_{k=2}^{m} \left(\left[\cap_{j=1}^{k-1} (A_j \cap B_j^c) \right] \cap B_k \right) \right\}, \tag{14}$$

where $A_k = \{S_r^{(k)} \in [a_r^{(k)}, N_k], S_t^{(k)} \in [a_t^{(k)}, N_k]\}$ and $B_k = \{S_r^{(k)} \in [b_r^{(k)}, N_k], S_t^{(k)} \in [b_t^{(k)}, N_k]\}$ with $N_k = \sum_{j=1}^{k} n_j$. Here, event $\cap_{j=1}^{k-1}(A_j \cap B_j^c)$ says the trial enters Stage k, while B_k says the treatment is established in Stage k.

The next result discusses the monotonicity of the probability function $P(R_m)$, the maximum Type I error rate, and the minimum power of the test R_m for testing (1).

Theorem 5.1. *Assume $\underline{X}^{(k)}$'s are independent for $k = 1, \ldots, m$ and each $\underline{X}^{(k)} \sim Multinomial(n_k, \underline{p})$ with $\underline{p} = (p_{11}, p_{10}, p_{01})$. Let $f(\underline{p}^*) = P(R_m)$ be a function of $\underline{p}^* = (p_r, p_t, p_{11})$. Then,*

(i) $f(\underline{p}^*)$ is nondecreasing in each of p_r and p_t when the other argument and p_{11} are fixed;

(ii) the maximum Type I error rate of the test R_m is

$$\max_{H_0} P(R_m) = \max_{H_0} f(\underline{p}^*) = \max \left\{ f(\underline{p}_{0r}^*), f(\underline{p}_{0t}^*) \right\};$$

(iii) the minimum power of the test R_m is

$$\min_{H_A} P(R_m) = \min_{H_A} f(\underline{p}^*) = \min_{p_{11} \in I_A} f(\underline{p}_A^*).$$

Furthermore, the expected total sample size, as a function of $\underline{p}^*$, is

$$EN(\underline{p}^*) = n_1 + \sum_{k=2}^{m} n_k P\left(\cap_{j=1}^{k-1}(A_j \cap B_j^c)\right).$$

Let EN_0 denote $\max\{EN(\underline{p}^*_{0r}), EN(\underline{p}^*_{0t})\}$ again. Then we search for the optimal m-stage design by determining two sequences of integer pairs $(a_r^{(k)}, b_r^{(k)})$ and $(a_t^{(k)}, b_t^{(k)})$ for $k = 1, \ldots, m$. The expected total sample size EN_0 is minimized while controlling the maximum Type I error rate no larger than α, the minimum power no smaller than $1 - \beta$, and the total sample size no larger than N^*. More computing power is needed because of the nature of computational complexity.

6 | DISCUSSION

In clinical trials, the assessments on the effectiveness and safety of a treatment are both of great interest. However, there are limited results on the construction of two-stage and multistage designs that consider the two aspects and also control the maximum Type I and II error rates. In phase II clinical trials, sample sizes are not large, so controlling the two error rates using exact tests is critical for making reliable statistical inferences. In this paper, we try to build a solid mathematical foundation of constructing optimal designs with control of the error rates.

When only the effectiveness of a treatment is under consideration, there are successful efforts to derive exact tests for (2). In this case, the distribution of the binary response depends on a single parameter p_r. Therefore, any test with a nondecreasing power function in p_r controls the two error rates over the null and alternative hypothesis spaces in (2) if the test controls the error rates at $p_r = p_{r_0}$ and $p_r = p_{r_1}$, the boundaries of the two spaces. Wang, Yin, and Zhang (2018) provided a solution for this problem in adaptive two-stage designs that allow a random n_2.

Conaway and Petroni (1995) first considered a version of (1) using multinomial data. They proposed to control the maximum probability of Type I error at a large level 0.3; however, they did not show the control of this error. Due to the multinomial data in Table 1, there are three parameters, p_r, p_t, and p_{11}, where p_{11} is a nuisance parameter. Different from the case of binomial data with a single parameter p_r, the boundaries of null and alternative spaces in (1) are indeed two-dimensional surfaces (these spaces are shown in Figure 1 without p_{11}) and no longer single points as in (2). For example, a point $(p_r, p_t) = (0.4, 0.4)$ is indeed a line segment of $(p_r, p_t, p_{11}) = (0.4, 0.4, p_{11})$ for $p_{11} \in [0, 0.4]$, while $(p_r, p_t) = (1, 1)$ is a single point $(p_r, p_t, p_{11}) = (1, 1, 1)$. Thus, it is not trivial to identify the parameter points with the maximum Type I error rate and the parameter point with the minimum power, because giving tests with a nondecreasing power function in each of p_r and p_t is not sufficient to control the two error rates.

In this paper, we propose a large family of tests in (5) that allow early termination due to either a successful or an unsuccessful treatment. Theorems 2.2 and 2.3 control the maximum probability of Type I error and the minimum power, respectively, by using a dimensionality reduction technique (see the proof of Theorem 2.2 and Lemma 6.3 in Appendix) to identify two parameter points, $\underline{p}^*_{0r}$, $\underline{p}^*_{0t}$ and one line $\underline{p}^*_A$. Hence, the mathematical foundation for exact tests in two-stage designs is built and extended to multistage designs. However, Jin (2007) deserved the credit of controlling the maximum Type I error rate for (6).

In construction of an optimal design Q^{opt}, we know how to find the maximum Type I error rate and the minimum power. For each fixed design Q, only three calculations are needed. However, for a general form Q in (4) with 12 quantities, it is not practical to use a single personal computer to find the optimal Q^{opt} by an exhausted search. One should use additional information to reduce the number of quantities in Q, shrink the search range on Q as in Theorem 3.1 and Step II in Section 4, or adopt more efficient programming skills. In order to reduce the expected total sample size, it would be of great interest to construct adaptive two-stage designs with consideration of both response and toxicity and establish the control on the maximum Type I error rate and the minimum power.

ACKNOWLEDGMENTS

The research of H.Y. and Z.Z. is partially supported by the National Natural Science Foundation of China (No. 11271039) and the Education Ministry Funds (No. 006000546615539). The authors thank two referees and the associate editor for their helpful comments and suggestions that substantially improved the paper.

CONFLICT OF INTEREST

The authors have declared no conflict of interest.

ORCID

Weizhen Wang https://orcid.org/0000-0001-7078-0485

REFERENCES

Banerjee, A., & Tsiatis, A. A. (2006). Adaptive two-stage designs in phase II clinical trials. *Statistics in Medicine*, *25*, 3382–3395.

Bryant, J., & Day, R. (1995). Incorporating toxicity considerations into the design of two-stage phase II clinical trials. *Biometrics*, *51*, 1372–1383.

Conaway, M. R., & Petroni, G. R. (1995). Bivariate sequential designs for phase II trials. *Biometrics*, *51*, 656–664.

Jin, H. (2007). Alternative designs of phase II trials considering response and toxicity. *Contemporary Clinical Trials*, *28*, 525–531.

Jin, H., & Wei, Z. (2012). A new adaptive design based on Simon's two-stage optimal design for phase II clinical trials. *Contemporary Clinical Trials*, *33*, 1255–1260.

Lin, Y., & Shih, W. J. (2004). Adaptive two-stage designs for single-arm phase IIA cancer clinical trials. *Biometrics*, *60*, 482–490.

Ray, H. E., & Rai, S. N. (2011). An evaluation of a Simon 2-Stage phase II clinical trial design incorporating toxicity monitoring. *Contemporary Clinical Trials*, *32*, 428–436.

Shan, G., Wilding, G. E., Hutson, A. D., & Gerstenberger, S. (2016a). Optimal adaptive two-stage designs for early phase II clinical trials. *Statistics in Medicine*, *35*, 1257–1266.

Shan, G., Zhang, H., Jiang, T., Peterson, H., Young, D., & Ma, C. (2016b). Exact p-values for Simon's two-stage designs in clinical trials. *Statistics in Biosciences*, *8*(2), 351–357.

Simon, R. (1989). Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials*, *10*, 1–10.

Song, J. X. (2015). A two-stage design with two co-primary endpoints. *Contemporary Clinical Trials Communications*, *1*, 2–4.

Stallard, N., Thall, P. F., & Whitehead, J. (1999). Decision theoretic designs for phase II clinical trials with multiple outcomes. *Biometrics*, *55*, 971–977.

Thall, P. F., Simon, R. M., & Estey, E. H. (1995). Bayesian sequential monitoring designs for single-arm clinical trials with multiple outcomes. *Statistics in Medicine*, *14*, 357–379.

Thall, P. F., & Sung, H. G. (1998). Some extensions and applications of a Bayesian strategy for monitoring multiple outcomes in clinical trials. *Statistics in Medicine*, *17*, 1563–1580.

Wang, W., Yin, H., & Zhang, Z. (2018). On exact inferences using binary data in two or multi-stage designs. Manuscript, Wright State University.

Wu, C., & Liu, A. (2007). An adaptive approach for bivariate phase II clinical trial designs. *Contemporary Clinical Trials*, *28*, 482–486.

Yu, J., Kepnerb, J. L., & Bundy, B. N. (2007). Exact power calculations for detecting hypotheses involving two correlated binary outcomes. *Statistics and Probability Letters*, *77*, 288–294.

SUPPORTING INFORMATION

Additional Supporting Information including source code to reproduce the results may be found online in the supporting information tab for this article.

How to cite this article: Yin H, Wang W, Zhang Z. On construction of single-arm two-stage designs with consideration of both response and toxicity. *Biometrical Journal*. 2019;61:1462–1476. https://doi.org/10.1002/bimj.201800320

APPENDIX: PROOFS

Proof of Theorem 2.1. The proofs for (i) and (ii) are similar, so, we only prove (i).

For (i), as in Table 1, p_r is an increasing function of p_{10} when p_t and p_{11} are fixed. Therefore, it suffices to show that $P(R)$ is nondecreasing in p_{10} for fixed p_{01} and p_{11}. Note that

(a) $P(R) = E[E(I_R|X_{11}, X_{01}, Y_{11}, Y_{01})]$ for an indicator $I_R = I_R(X_{11}, X_{10}, X_{01}, Y_{11}, Y_{10}, Y_{01})$;

(b) the conditional distributions of X_{10} and Y_{10} for given $(X_{11}, X_{01}, Y_{11}, Y_{01})$ are

$$\begin{aligned} X_{10}\Big|X_{11}, X_{01} &\sim Binomial\left(n_1 - x_{11} - x_{01}, \frac{p_{10}}{1 - p_{11} - p_{01}}\right), \\ Y_{10}\Big|Y_{11}, Y_{01} &\sim Binomial\left(n_2 - y_{11} - y_{01}, \frac{p_{10}}{1 - p_{11} - p_{01}}\right); \end{aligned} \tag{A1}$$

(c) the joint distribution of $(X_{11}, X_{01}, Y_{11}, Y_{01})$ does not involve p_{10}.

We only need to show $E(I_R|x_{11}, x_{01}, y_{11}, y_{01})$ is nondecreasing in p_{10}, which is based on Lemma 6.2 below (however, Lemma 6.2 is based on Lemma 6.1). Let $h(p_{10}) = E(I_R|x_{11}, x_{01}, y_{11}, y_{01})$. By part (a) of Lemma 6.2, $h(p_{10}) = f_1(p_{10}, p_{10})$, where f_1 is introduced in Lemma 6.2. For $p_{10} < p'_{10}$, we have

$$h(p_{10}) = f_1(p_{10}, p_{10}) \le f_1(p_{10}, p'_{10}) \le f_1(p'_{10}, p'_{10}) = h(p'_{10})$$

following part (b) of Lemma 6.2. The proof for (i) is complete.

Lemma 6.1. *For fixed $x_{11}, x_{01}, y_{11}, y_{01}$, consider the indicator function I_R as a function of (X_{10}, Y_{10}), denoted by $I_R(X_{10}, Y_{10})$. Then, $I_R(X_{10}, Y_{10})$ is nondecreasing in any of its arguments when the other argument is fixed.*

Proof of Lemma 6.1. First, it is easy to see that $I_R(x_{10}, Y_{10})$ is nondecreasing in Y_{10} for a fixed x_{10}.

Second, for a fixed y_{10}, rewrite $I_R(X_{10}, y_{10})$ as

$$\begin{aligned} I_R(X_{10}, y_{10}) = {} & I_{\{X_{10} \ge b_r - x_{11}, x_{11}+x_{01} \ge b_t\}} \\ & + I_{\{\max\{a_r - x_{11}, c_{r_1} - x_{11} - y_{11} - y_{10}\} \le X_{10} < b_r - x_{11}, x_{11}+x_{01} \ge b_t, x_{11}+x_{01}+y_{11}+y_{01} \ge c_{t_1}\}} \\ & + I_{\{X_{10} \ge \max\{b_r - x_{11}, c_{r_2} - x_{11} - y_{11} - y_{10}\}, a_t \le x_{11}+x_{01} < b_t, x_{11}+x_{01}+y_{11}+y_{01} \ge c_{t_2}\}} \\ & + I_{\{\max\{a_r - x_{11}, c_{r_3} - x_{11} - y_{11} - y_{10}\} \le X_{10} < b_r - x_{11}, a_t \le x_{11}+x_{01} < b_t, x_{11}+x_{01}+y_{11}+y_{01} \ge c_{t_3}\}}. \end{aligned} \tag{A2}$$

So, (a) the sum of the first two lines on the right-hand side of (A2) is nondecreasing in X_{10}; (b) the same occurs for the sum of the last two lines. These imply that $I_R(X_{10}, y_{10})$ is nondecreasing in X_{10}.

Lemma 6.2. *Suppose*

$$X'_{10} \Big| X_{11}, X_{01} \sim \textit{Binomial}\left(n_1 - x_{11} - x_{01}, \frac{p^{(1)}_{10}}{1 - p_{11} - p_{01}}\right),$$
$$Y'_{10} \Big| Y_{11}, Y_{01} \sim \textit{Binomial}\left(n_2 - y_{11} - y_{01}, \frac{p^{(2)}_{10}}{1 - p_{11} - p_{01}}\right),$$

for $0 \le p^{(1)}_{10}, p^{(2)}_{10} \le 1 - p_{11} - p_{01}$ and they are independent. Let $f_1(p^{(1)}_{10}, p^{(2)}_{10}) = EI_R(X'_{10}, Y'_{10})$ be a function of $p^{(1)}_{10}$ and $p^{(2)}_{10}$. Then,

(a) $f_1(p_{10}, p_{10}) = EI_R(X_{10}, Y_{10})$;
(b) $f_1(p^{(1)}_{10}, p^{(2)}_{10})$ is nondecreasing in any of the $p^{(i)}_{10}$'s for $i = 1, 2$ when the other $p^{(i)}_{10}$ is fixed.

Proof of Lemma 6.2. Part (a) follows (A1). For part (b), fix $p^{(2)}_{10}$ first. We rewrite $I_R(X'_{10}, Y'_{10})$ as $I_R(X'_{10}, Y'_{10}) = I_{\{X'_{10} \ge f_2(Y'_{10})\}}$ for some function f_2 since, by Lemma 6.1, $I_R(X'_{10}, Y'_{10})$ is nondecreasing in X'_{10} when Y'_{10} is fixed. Note the probability $P(X \ge a)$ is a nondecreasing function of p for any value a when $X \sim \textit{Binomial}(n, p)$. Thus, combining with the independence of X'_{10} and Y'_{10}, $f_1(p^{(1)}_{10}, p^{(2)}_{10}) = E(I_R(X'_{10}, Y'_{10})) = P(X'_{10} \ge f_2(Y'_{10}))$ is nondecreasing in $p^{(1)}_{10}$. Similarly, $f_1(p^{(1)}_{10}, p^{(2)}_{10})$ is nondecreasing in $p^{(2)}_{10}$.

Proof of Theorem 2.2. Write $H_0 = H_{r_0} \cup H_{t_0}$, where $H_{r_0} = \{p_r \le p_{r_0}\}$ and $H_{t_0} = \{p_t \le p_{t_0}\}$. It is clear that $\sup_{H_{r_0}} P(R) \ge P(R|p_r = p_{r_0}, p_t = 1) = f(p_{r_0}, 1, p_{r_0}) = f(\underline{p}^*_{0r})$. Next, we prove $\sup_{H_{r_0}} P(R) \le f(\underline{p}^*_{0r})$. Note $R \subset \{(X_r \ge b_r\} \cup \{a_r \le X_r < b_r, X_r + Y_r \ge c_{r_1})\}$. Following Theorem 1 in Wang et al. (2018), where they focus on two-stage designs when only a univariate variable, the response, is observed but n_2 depends on X_r in Stage 1, we obtain

$$\sup_{H_{r_0}} P(R) \le \sup_{H_{r_0}} [P(X_r \ge b_r) + P(a_r \le X_r < b_r, X_r + Y_r \ge c_{r_1})] = f(\underline{p}^*_{0r}).$$

Similarly, we can show $\sup_{H_{t_0}} P(R) = f(\underline{p}^*_{0t})$ and conclude Theorem 2.2.

Lemma 6.3. *Consider $g(\underline{p}) = P(R)$ for $\underline{p} = (p_{11}, p_{10}, p_{01})$. For a constant c, $g(\underline{p})$ is nondecreasing in p_{11} when any of the following two conditions is true:*

(i) if $p_{11} + p_{10} = c$ and $p_{01} = 0$ (i.e., $X_{01} = 0 = Y_{01}$);

(ii) if $p_{11} + p_{01} = c$ and $p_{10} = 0$ (i.e., $X_{10} = 0 = Y_{10}$).

Proof of Lemma 6.3. The proofs for (i) and (ii) are similar. We only prove (i) here.

Because $p_{11} + p_{10} = c$ and $p_{01} = 0$, $(X_{10}, X_{01}) = (X_r - X_{11}, 0)$ and $(Y_{10}, Y_{01}) = (Y_r - Y_{11}, 0)$. Then the indicator function I_R, which originally is equal to $I_R(X_{11}, X_{10}, X_{01}, Y_{11}, Y_{10}, Y_{01})$, can be rewritten as $I_R(X_{11}, X_r, Y_{11}, Y_r)$. Furthermore, for given (X_r, Y_r), I_R is a function of X_{11} and Y_{11}, denoted by $I_R(X_{11}, Y_{11})$. Note $g(\underline{p}) = E[E(I_R|X_r, Y_r)]$. The conditional distributions of $X_{11}|X_r$ and $Y_{11}|Y_r$ are

$$X_{11}\Big|X_r \sim \textit{Binomial}\left(x_r, \frac{p_{11}}{c}\right), Y_{11}\Big|Y_r \sim \textit{Binomial}\left(y_r, \frac{p_{11}}{c}\right), \text{respectively.} \tag{A3}$$

Similar to Theorem 2.1, one can show that $g(\underline{p})$ is nondecreasing in p_{11}. The proof for i) is complete.

Proof of Theorem 2.3. For any point $\underline{p}^* = (p_r, p_t, p_{11})$ in H_A, $p_r \geq p_{r_1}$, $p_t \geq p_{t_1}$ and $p_{11} \geq a$ with $a = \max\{0, p_r + p_t - 1\}$. Due to Part (i) of Theorem 2.1, the power for the test R in (5), $f(\underline{p}^*)$, satisfies

$$f(\underline{p}^*) = f(p_r, p_t, p_{11}) \geq f(\max\{p_{r_1}, p_{11}\}, p_t, p_{11}). \tag{A4}$$

(a) If $p_{r_1} \geq p_{11}$, following Part (ii) of Theorem 2.1,

$$\text{the last quantity in (A4)} = f(p_{r_1}, p_t, p_{11}) \geq f(p_{r_1}, \max\{p_{t_1}, p_{11}\}, p_{11}) \overset{\text{denoted by}}{=} A.$$

If $p_{t_1} \geq p_{11}$, then $A = f(p_{r_1}, p_{t_1}, p_{11})$; if $p_{t_1} < p_{11}$, then $A = f(p_{r_1}, p_{11}, p_{11}) \geq f(p_{r_1}, p_{t_1}, p_{t_1})$ due to Part (i) of Lemma 6.3.

(b) If $p_{r_1} < p_{11}$, following Part (ii) of Theorem 2.1,

$$\text{the last quantity in (A4)} = f(p_{11}, p_t, p_{11}) \geq f(p_{11}, \max\{p_{t_1}, p_{11}\}, p_{11}) \overset{\text{denoted by}}{=} B.$$

(b-1) If $p_{t_1} \geq p_{11}$, $B = f(p_{11}, p_{t_1}, p_{11}) = g(p_{11}, 0, p_{t_1} - p_{11}) \geq g(p_{r_1}, 0, p_{t_1} - p_{r_1}) = f(p_{r_1}, p_{t_1}, p_{r_1})$ due to Part (ii) of Lemma 6.3.

(b-2) If $p_{t_1} < p_{11}$, following Theorem 1 in Wang et al. (2018),

$$B = f(p_{11}, p_{11}, p_{11}) \geq f(\max\{p_{r_1}, p_{t_1}\}, \max\{p_{r_1}, p_{t_1}\}, \max\{p_{r_1}, p_{t_1}\}) \overset{\text{denoted by}}{=} B_2.$$

If $p_{r_1} \geq p_{t_1}$, following Part (i) of Lemma 6.3, $B_2 = g(p_{r_1}, 0, 0) \geq g(p_{t_1}, p_{r_1} - p_{t_1}, 0) = f(p_{r_1}, p_{t_1}, p_{t_1})$; if $p_{r_1} < p_{t_1}$, $B_2 = g(p_{t_1}, 0, 0) \geq g(p_{r_1}, 0, p_{t_1} - p_{r_1}) = f(p_{r_1}, p_{t_1}, p_{r_1})$ due to Part (ii) of Lemma 6.3.

Lemma 6.4. *Let $\mathcal{Q}_{B0}$ be the collection of Q_B satisfying the first equation in* (11) *and $P(R_B|\underline{p}^* = \underline{p}^*_{A0}) \geq 1 - \beta$, and let Q_{B0}^{opt} be the design in $\mathcal{Q}_{B0}$ that minimizes EN_0. Then $Q_{B0}^{opt} = Q_B^{opt}$ if* (13) *holds.*

Proof of Lemma 6.4. Let $\mathcal{Q}_B$ be the collection of all Q_B satisfying (11). Then, Q_B^{opt} minimizes EN_0 in $\mathcal{Q}_B$, and $\mathcal{Q}_B \subset \mathcal{Q}_{B0}$. If (13) holds, then $Q_{B0}^{opt} \in \mathcal{Q}_B$, indicating $Q_{B0}^{opt} = Q_B^{opt}$.

Proof of Theorem 3.1. We only show that $g_{R_B}(a_r, a_t, b_r, b_t, c_r, c_t)$ is nonincreasing in b_r as the proofs for the other claims in Theorem 3.1 are either similar or trivial. Note

$$\begin{aligned}
&g_{R_B}(a_r, a_t, b_r + 1, b_t, c_r, c_t) - g_{R_B}(a_r, a_t, b_r, b_t, c_r, c_t) \\
&= P(X_r = b_r, X_t \geq b_t, X_r + Y_r \geq c_r) - P(X_r = b_r, X_t \geq b_t) \\
&\quad + P(X_r = b_r, a_t \leq X_t < b_t, X_r + Y_r \geq c_r, X_t + Y_t \geq c_t) \\
&\quad - P(X_r = b_r, a_t \leq X_t < b_t, X_t + Y_t \geq c_t) \leq 0.
\end{aligned}$$

Proof of Theorem 5.1. The proof is similar to the proofs of Theorems 2.1 – 2.3, and is omitted.