Single-arm phase II three-outcome designs with handling of over-running/under-running

Wenchuan Guo, Jianan Hui & Bob Zhong

STATISTICAL THEORY AND RELATED FIELDS
2023, VOL. 7, NO. 4, 276–286
https://doi.org/10.1080/24754269.2023.2189348

Taylor & Francis
Taylor & Francis Group

OPEN ACCESS   Check for updates

Single-arm phase II three-outcome designs with handling of over-running/under-running

Wenchuan Guo[a], Jianan Hui[b] and Bob Zhong[c]

[a]Seagen Inc., Bothell, WA, USA; [b]Servier Pharmaceuticals, Boston, MA, USA; [c]Regeneron Pharmaceuticals, Tarrytown, NY, USA

ABSTRACT

Phase II clinical trials are commonly conducted as pilot studies to evaluate the efficacy and safety of the investigational drug in the targeted patient population with the disease or condition to be treated or prevented. When designing such a trial considering efficacy conclusions, people naturally think as follows: if efficacy evidence is very strong, a go decision should be made; if efficacy evidence is very weak, a no-go decision should be made; if the efficacy evidence is neither strong nor weak, no decision can be made (inconclusive). The designs presented in this paper match this natural thinking process with go/no-go/inconclusive outcomes. Both two-/three-stage designs are developed with three outcomes. Additionally, a general approach based on conditional error function is implemented such that new decision boundaries can be calculated to handle mid-course sample size change which results in either 'over-running' or 'under-running' and ensure the control of overall type I error. A free open-source R package `tsdf` that calculates the proposed two-/three-stage designs is available on `CRAN`.

ARTICLE HISTORY
Received 6 September 2022
Revised 2 March 2023
Accepted 6 March 2023

KEYWORDS
Phase II trials; two-stage and three-stage designs; go/no-go/inconclusive outcomes; asymmetric alpha; alpha spending; optimal designs; over-running and under-running

1. Introduction

Phase II clinical trials are commonly conducted as pilot studies to evaluate the efficacy and safety of the investigational drug in the targeted patient population of the disease or condition to be treated or prevented. The initial assessment of efficacy plays an important role in determining if a compound should be further studied in the targeted patient population. When designing such a trial considering efficacy conclusions, people naturally think as follows: if efficacy evidence is very strong, a go decision should be made; if efficacy evidence is none or very weak, a no-go decision should be made; if the efficacy evidence is neither very strong nor very weak, no clear decision can be made (inconclusive). This decision process is essential in proof-of-concept (POC) studies and this is fundamentally different from Phase III confirmatory studies. Note that this decision process is quite common in real-world practice. For example, in one oncology study (the actual tumour type and compound are not revealed here since the compound is still under development), the decision tree is as follows: if the response rate is above 50%, a phase 3 study will be launched with a single agent under study as the active treatment against an active control; if the response rate is below 30%, no further studies will be conducted on the compound for the cancer under study; if the response rate is between 30% and 50%, a study of the agent in combination with an existing treatment will be conducted. The question is how to design a trial that matches this natural decision process and supports it statistically.

The topic has been discussed in B. Zhong (2012) and W. Zhong and Zhong (2013) and they proposed a design to accommodate such a natural and practical thinking process that can provide a more realistic basis for decision making in Phase 2 trials comparing to the most commonly used Simon's two-stage design. The design is outlined as follows. First, the minimal effective response rate, p_c, is selected. Then the hypotheses are set up as: $H_0 : p = p_c$, vs. $H_1 : p \neq p_c$, where if data supports hypothesis conclusion $p < p_c$ then a no-go decision should be made, and if data supports hypothesis conclusion $p > p_c$ then a go decision should be made. If data supports neither hypothesis conclusion $p < p_c$ nor hypothesis conclusion $p > p_c$ then neither a go nor a no-go decision can be made, resulting in an inconclusive outcome. Similar designs with three outcomes under different hypothesis settings were developed to overcome the issue that the union of null and alternative hypotheses can't cover the whole parameter spaces (Hong & Wang, 2007; Sargent et al., 2001). In practice, Zhong's two-stage designs have some limitations, which motivated this paper. The objective of our paper is to modify and extend Zhong's two-stage designs to address the following. (1) Allow early stopping for efficacy. If the intervention proves to be superior, the study can be stopped early to maximize the number of patients who will benefit from the intervention, to save time and to reduce costs in drug development. (2) Apply alpha spending method (Demets & Lan, 1994) on two-/three-stage designs so that type I

CONTACT Bob Zhong ✉ chongxin.zhong@regeneron.com 🖃 Regeneron Pharmaceuticals, Tarrytown, NY 10591, USA

error allocated at each stage will be controlled. (3) Extend the design where an interval (p_l, p_u) is utilized instead of a single point p_c to accommodate the uncertainty in specifying p_c in practice. Specifically, if the true response rate of the test treatment is p and if $p < p_l$ then the test treatment is clearly not effective (no-go) or if $p > p_u$ then the test treatment is clearly effective (go). (4) Allow for mid-course sample size change. In classical group sequential designs, the sample sizes for the interim and final analyses need to be pre-specified and followed strictly throughout the trial to control the overall type I error rate. In practice, however, the actual sample sizes often deviate from the planned ones, such as trials conducted in multi-centres. These situations are referred to 'over-running' and 'under-running' (Whitehead, 1992). To this end, we derive new decision boundaries based on the conditional error function and ensure strictly control of the overall type I error rate under these 'over-running' or 'under-running' situations. Various procedures are proposed to deal with unplanned sample size at interim or final stage (Koyama & Chen, 2008; Li et al., 2002; Shan & Chen, 2018). However, these existing methods either based on Simon's two-stage design, not providing flexible stopping rule for early stopping for efficacy, or don't control type 1 error rate at each stage.

The remainder of the paper is organized as follows. Section 2 is devoted to explain how to calculate two- and three-stage designs. Section 3 explains the handling of 'over-running' and 'under-running' using conditional error function. Section 4 introduces an R package `tsdf` that implements the proposed two-/three-stage designs. We provide some examples in Section 5 to demonstrate our proposed designs and the pragmatic feature of handling the 'over-running' and 'under-running'. Discussions are given in Section 6.

2. Two- and three-stage designs

The key of Zhong's two-stage design (B. Zhong, 2012) is to correctly specify the minimal effective response rate, p_c. Intuitively, any response rate below p_c (the minimal effective response rate) is considered ineffective hence does not warrant further development. In contrast, any response rate above p_c is considered effective hence may warrant further development. It can be seen that the selection of the minimal effective response rate (threshold) is critical. General guidelines are given below on how to select the threshold in single-arm trials. There are three scenarios in practice. 1. When a standard of care for the disease and population under investigation is available, a reliable point estimate of the response rate of the standard care is commonly available. In this scenario, if the test treatment can be developed as 'similar' to the standard of care, then one of the choices is to select the point estimate of the standard of care as the minimal effective response rate. This is based on the fact that the standard of care is an effective therapy. If the test treatment is intended to be developed as a treatment superior over the standard treatment, then a rate equal to or higher than the point estimate of the standard care can be chosen according to clinical and statistical judgment. For oncology studies where the objective response rate is the endpoint, the minimal effective response rate may be chosen as 5% to 10% above the point estimate. 2. When a standard of care is not available but several treatments are available, then the selection of the minimal response rate depends on the choice of the control of future randomized registration studies. If one of the available treatments will be the active control, then that treatment can serve as the 'standard of care' for the purpose of determining of the minimal effective response rate. The patient population should mimic the population of the control as well. If the objective is to beat all of them, then a possible choice is to use the highest point estimate of the response rates among all available treatments. 3. When there is no treatment available, the minimal effective response rate can be set based on clinical and statistical judgment. Historical data under best supportive care is commonly used to support such a choice.

Once the minimal effective response rate p_c is chosen, the hypotheses are set as

$$H_0 : p = p_c \quad \text{vs.} \quad H_1 : p \neq p_c. \tag{1}$$

Note that the alternative can be decomposed as two parts:

$$H_1^- : p < p_c \quad \text{and} \quad H_1^+ : p > p_c.$$

There are three possible conclusions for the above hypothesis test: do not reject null hypothesis, reject null and conclude H_1^+, and reject null and conclude H_1^-. It means we may conclude that the response rate is equal, higher or lower than the minimal effective response rate. This hypothesis test also can be generalized to the case that the minimal effective response rate is not a single value but a pre-specified interval. The hypothesis becomes

$$H_0 : p \in [p_l, p_u] \quad \text{vs.} \quad H_1 : p \notin [p_l, p_u], \tag{2}$$

where $p_u > p_l$. Similarly, we decompose the alternative as

$$H_1^- : p < p_l \quad \text{and} \quad H_1^+ : p > p_u.$$

The interval setup is motivated by situations where a single point cannot be confidently and accurately determined. For example, when a standard of care exists but there is more than one reliable and large trial that yields different estimates of the response of the standard of care. Another example is when there is no available treatment and clinical judgment is expressed in an interval, the response rate below 15% is clearly not of interest and the response rate above 25% is clearly of interest for further development. In these situations, the threshold can be set as an interval $[p_l, p_u]$. That is, any response rate below p_l is considered ineffective hence does not warrant further development; any response rate above p_u is considered effective hence may warrant further development. The hypothesis test in (2) is equivalent to test in (1) when $p_l = p_u = p_c$, so we proceed with (2) hereafter.

Denote x_i as the cumulative number of responders among $n_1 + n_2 + \cdots + n_i$ at stage i. The corresponding left-side critical values are r_i's and right-side critical values are s_i's. At each stage, one of the following decisions will be made.

- If $x_i \leq r_i$, conclude H_1^- and stop the trial for inefficacy.
- If $x_i > s_i$, conclude H_1^+ and stop the trial for efficacy.
- If $r_i < x_i \leq s_i$, proceed to next stage and treat an additional n_{i+1} subject.

r_i and s_i are determined by significance levels, i.e., left-side and right-side type I errors. Denote the overall left-side type I error as α_1, right-side type I error as α_2, and the type II error as β. We use the α-spending function to distribute the overall type I error over two/three stages, which prevents the case that most α are spent in the early stages. We denote the cumulative left-side type I errors at stage i as α_{1i}'s ($\alpha_{11} \leq \alpha_{12}$) and the cumulative right-side type I errors as α_{2i}'s ($\alpha_{21} \leq \alpha_{22}$). We have the following constraints:

- if $p = p_l$, the probability of concluding H_1^- should not exceed α_{1i} (left-side type I error) ;
- if $p = p_u$, the probability of concluding H_1^+ should not exceed α_{2i} (right-side type I error);
- if the expected response rate is p_e ($> p_u$), then the probability of not concluding H_1^+ should not exceed β.

Before we provide details of two-stage designs and three-stage designs in the following sections, how error constraints affect the trial design will be discussed. The type I error is the probability of rejecting the true null hypothesis. For left-side, high type I error means that it's more likely to conclude H_1^-, i.e., it's easier to terminate the trial for inefficacy. Thus high left-side type I error designs lead to a higher chance of terminating the trial for inefficacy. Right side is the opposite: low right-side type I error is more conservative as it is harder to reject the null hypothesis, which leads to declaring efficacy outcome. Investigators can choose a suitable design by giving specific left-side, right-side type I errors and type II error, respectively.

We describe two-stage designs in Section 2.1 and three-stage designs in Section 2.2.

2.1. Two-stage designs

The two-stage design setup is: n_1 patients are treated in the first stage. If the trial continues to the second stage, additional n_2 patients are treated. Recall that x_i is the total cumulative number of responders until stage i. The procedure is as follows ($r_i \leq s_i, r_1 \leq r_2, s_1 \leq s_2, r_i \leq \sum_{k=1}^{i} n_k, s_i \leq \sum_{k=1}^{i} n_k$).

(1) Stage 1: treat n_1 patients
  - If $x_1 \leq r_1$, terminate the trial and conclude H_1^-.
  - If $x_1 > s_1$, terminate the trial and conclude H_1^+.
  - If $r_1 < x_1 \leq s_1$, continue the trial and go to stage 2.

(2) Stage 2: treat additional n_2 patients
  - If $x_2 \leq r_2$, terminate the trial and conclude H_1^-.
  - If $x_2 > s_2$, terminate the trial and conclude H_1^+.
  - If $r_2 < x_2 \leq s_2$, terminate the trial and conclude 'data does not contradict to null hypothesis'.

Denote the binomial cumulative density function as $B(\cdot; n, p)$ and probability function as $b(\cdot, n, p)$, where n is the number of Bernoulli trials and p is the probability of success. Let's calculate the conditional probabilities. If the true response rate is p and given r_1, n_1, then the probability of concluding H_1^- at the first stage is

$$L_1(p) = B(r_1, n_1, p) \tag{3}$$

and at the second stage, given r_2, n_2, is

$$L_2(p) = \sum_{t_1=r_1+1}^{s_1} b(t_1, n_1, p)B(r_2 - t_1, n_2, p). \tag{4}$$

Similarly, the probabilities of concluding H_1^+ at stage 1 and 2 are

$$R_1(p) = 1 - B(s_1, n_1, p) \quad \text{and} \quad R_2(p) = \sum_{t_1=r_1+1}^{s_1} b(t_1, n_1, p)[1 - B(s_2 - t_1, n_2, p)].$$

Lower and upper type I errors α_1, α_2 are spent following an error spending method. For any chosen error spending function, error rate $\alpha_{i1} \le \alpha_{i2} = \alpha_i$ allowed at each stage can be calculated. Therefore, r_i, s_i satisfy the following type I error constraints:

$$L_1(p_l) \le \alpha_{11}, \quad L_1(p_l) + L_2(p_l) \le \alpha_{12} = \alpha_1 \tag{5}$$

and

$$R_1(p_u) \le \alpha_{21}, \quad R_1(p_u) + R_2(p_u) \le \alpha_{22} = \alpha_2. \tag{6}$$

The type II error constraint is

$$\sum_{i=1}^{2} R_i(p_e) \ge 1 - \beta. \tag{7}$$

We search all combinations of (r_i, s_i, n_i) satisfying the type I error and type II error constraints. The focus is to pursue designs that have the closest errors to the desired left-side and right-side type I errors under the condition that $\le \alpha_{i1}$ at stage 1 and $\le \alpha_{i2} = \alpha_i, i = 1, 2$ at stage 2 from a chosen α-spending function and the minimal sample size n with many choices of n_1 under the expected response rate. In addition, it is also desirable to maximize the power for a given n. With the conditions (5), (6), and (7) outlined, there are usually many designs with combinations of (r_i, s_i, n_i) that satisfy type I and II error constraints. Therefore, additional selection criteria are needed to choose designs that satisfy practical considerations. Given n_1, n_2, all possible combinations of $(r_i, s_i, i \le 2)$ satisfying the constraints are outputted into a matrix in R. The designs are then sorted in descending order by left-side, right-side type I errors and power. The first design is then chosen. For given n_1, n_2, this chosen design has the closest type I errors to α_{11}, α_{21}, α_{12}, and α_{22} and the maximum of power. We increment the total sample n from 2 until predetermined distinct choices of n_1 are found. Denote the smallest possible n_1 as n_{1s} and the largest as n_{1l}. Then all integers between n_{1s} and n_{1l} can be used as the first stage sample size. In practice, this means once the total sample is fixed, the stage 1 sample size can vary within a range instead being fixed, which makes trial conduct easier to manage.

With so many choices of stage 1 sample size as described above, a natural question is where we should set our target stage 1 sample size. One possible choice is to set the target sample size at which the expected sample size is minimized under null hypothesis. That is, we should target to have the design that minimizes $n_1 + n_2 \times (1 - B(r_1, n_1, p_l))$ (without early stopping for efficacy) or $n_1 + n_2 \times (B(s_1, n_1, p_u) - B(r_1, n_1, p_l))$ (with early stopping for efficacy). The chosen design is called optimal design.

2.2. Three-stage designs

Three-stage design is the extension of two-stage design where we treat additional n_3 patients if the decision is to continue to the next stage at the end of stage 2. Thus the sample size is at most $n_1 + n_2 + n_3 = n$. The complete three-stage design is as follows.

(1) Stage 1: treat n_1 patients
- If $x_1 \le r_1$, terminate the trial and conclude H_1^-.
- If $x_1 > s_1$, terminate the trial and conclude H_1^+.
- If $r_1 < x_1 \le s_1$, go to stage 2.

(2) Stage 2: treat additional n_2 patients
- If $x_2 \le r_2$, terminate the trial and conclude H_1^-.
- If $x_2 > s_2$, terminate the trial and conclude H_1^+.
- If $r_2 < x_2 \le s_2$, go to Stage 3.

(3) Stage 3: treat additional n_3 patients
- If $x_3 \leq r_3$, terminate the trial and conclude H_1^-.
- If $x_3 > s_3$, terminate the trial and conclude H_1^+.
- If $r_3 < x_3 \leq s_3$, terminate the trial and conclude 'data does not contradict to null hypothesis '.

Then we only need to calculate the conditional probabilities at the third stage in addition to the first two stages that have been calculated in previous subsection. Given $n_1, n_2, r_1, r_2, r_3, s_1, s_2$, the probability of concluding H_1^- at the third stage is

$$L_3(p) = \sum_{t_1=r_1+1}^{s_1} \sum_{t_2=r_2-t_1+1}^{s_2-t_1} b(t_1, n_1, p) b(t_2, n_2, p) B(r_3 - t_1 - t_2, n_3, p), \tag{8}$$

and concluding H_1^+ at the third stage is

$$R_3(p) = \sum_{t_1=r_1+1}^{s_1} \sum_{t_2=r_2-t_1+1}^{s_2-t_1} b(t_1, n_1, p) b(t_2, n_2, p)[1 - B(s_3 - t_1 - t_2, n_3, p)].$$

Combining (8) with the constrains on the first two stages (3), (4), r_i, n_i satisfy the following constraints:

$$\sum_{k=1}^{i} L_k(p_l) \leq \alpha_{1i} \quad \text{and} \quad \sum_{k=1}^{i} R_k(p_u) \leq \alpha_{2i},$$

for $i = 1, 2, 3$ and

$$\sum_{i=1}^{3} R_i(p_e) \geq 1 - \beta.$$

The optimal design is chosen as described in the end of Section 2.1.

3. Adjustment for over-running and under-running

To assure the control of type I error rate, the sample size at each stage need to be specified in the protocol and adhered strictly during study conduct in sequential clinical trial. However, a considerably amount of uncertainty is usually expected during the planning stage of a study and such strong restrictions of the sample sizes imposed by the common two-/three-stage designs can make study conduct very challenging. For example, stopping the recruitment exactly after certain number of patients have been accrued can be difficult especially in a multi-centre trial. Patients may have been screened and it is unethical to withhold treatment and exclude such patients from the study. As a result, over-running may occur. Moreover, even if we have recruited the exact number of patients as pre-specified in the protocol, it is possible that not all the patients are evaluable in regard to the primary endpoint, leading to a situation we then refer as 'under-running' of the study. In all the aforementioned cases, there is a violation of the predetermined sample sizes and the control of type I error is no longer guaranteed.

In general, there are three possible scenarios: (1) sample size is only modified at interim analysis; (2) sample size is only modified at final analysis; (3) sample size is modified at both interim analysis and final analysis. For the first scenario, recall that in our proposed two-/three-stage designs, a range of the designs satisfying the criterion are generated. In other words, our proposed design allows change of the sample size n_1 in the first stage for two-stage design or (n_1, n_2) in the first two stages in the three-stage design as long as the new sample size n_1^* or (n_1^*, n_2^*) is within the range and the total sample size remains the same as the initial design while controlling the type I error. For the third scenario, it is easy to locate a design in our list of proposed designs to match the modified sample size at interim. With that being said, the third scenario can be reduced to the second scenario. Therefore, we only need to consider the cases where the sample size at the final analysis is modified and discuss possible remedies for mid-course modifications of sample size while strictly controlling the overall type I error rate.

We apply the conditional error function approach in Englert and Kieser (2012, 2015) that allows for arbitrary modifications of sample size based on the results of the interim analysis or external information while controlling for overall type I error. The concept of conditional error function (Proschan & Hunsberger, 1995) was introduced as a method to test the null hypothesis within a two-stage design and allow for data dependent modifications of the sample sizes after the first stage. The conditional significance level of the second stage depends on the outcome of the first stage. Simultaneously, conditioning on the results from interim analysis, the rejection region of the final decision is invariant to the mid-course modifications such that the unconditional overall Type I error is controlled. The flexibility nature of such designs is referred as 'conditional invariance principle' (Brannath et al., 2007).

For single-arm oncology trials with the binary endpoint, the conditional error function is the type I error rate used at the final stage given the number of responses observed in the previous stages. We first apply the conditional error function approach to the proposed two-stage design. Assume the sample size modification happens at the second stage. After completion of the first stage, the trial proceeds to the second stage and the second stage sample size may be changed from n_2 to n_2^* so that we need to find new boundaries r_2^* and s_2^* to assure that overall type I error is controlled. Recall that our hypothesis is $H_0 : p \in [p_l, p_u]$ vs. $H_1 : p \notin [p_l, p_u]$ and the null hypothesis can be written as $H_1^- : p < p_l$ and $H_1^+ : p > p_u$. We allocate different type I error α_1 and α_2 on left side and right side, respectively. Denote the number of responses at the first stage as t_1. Then the conditional function for left side test H_1^- is

$$f_l(t_1) = \begin{cases} 0, & \text{if } t_1 \le r_1, \\ B(r_2^* - t_1, n_2^*, p_l), & \text{if } r_1 < t_1 \le s_1\ , \\ 1, & \text{if } t_1 > s_1, \end{cases}$$

and, for right side H_1^+,

$$f_r(t_1) = \begin{cases} 0, & \text{if } t_1 \le r_1, \\ 1 - B(s_2^* - t_1, n_2^*, p_u), & \text{if } r_1 < t_1 \le s_1, \\ 1, & \text{if } t_1 > s_1. \end{cases}$$

The conditional error functions can be considered as the conditional significance level that can be used at final analysis. The trial is terminated early for efficacy or inefficacy when the conditional function is equal to 0 and 1, respectively. We then have the left-side type I error

$$\alpha_1^* = \sum_{t_1=0}^{n_1} f_l(t_1, p_l) \cdot b(t_1, n_1, p_l),$$

and the right-side type I error

$$\alpha_2^* = \sum_{t_1=0}^{n_1} f_r(t_1, p_u) \cdot b(t_1, n_1, p_u).$$

The boundaries r_2^* and s_2^* are selected such that the actual type I error α_i^* is at most $\alpha_i, i = 1, 2$ under the null hypothesis. In all possible combinations of r_2^* and s_2^*, the one has the closest type I errors to the desired significance level and the smallest type II error is selected.

For three-stage designs, the third stage sample size may be changed from n_3 to n_3^* after completion of stage 1 and stage 2. We need to obtain new boundaries s_3^* and r_3^* to control overall type I error. Similarly, we obtain the right-side type I error α_2^{**}

$$\alpha_2^{**} = \sum_{t_1=0}^{n_1} \sum_{t_2=0}^{n_2} f_r(t_1, t_2) \cdot b(t_1, n_1, p_u) \cdot b(t_2, n_2, p_u),$$

where

$$f_r(t_1, t_2) = \begin{cases} 0, & \text{if } t_1 + t_2 \le r_2 \text{ or } t_1 \le r_1, \\ 1, & \text{if } t_1 + t_2 > s_2 \text{ or } t_1 > s_1, \\ 1 - B(s_3^* - t_1 - t_2, n_3^*, p_u), & \text{otherwise,} \end{cases}$$

and the left-side type I error α_1^{**}

$$\alpha_1^{**} = \sum_{t_1=0}^{n_1} \sum_{t_2=0}^{n_2} f_l(t_1, t_2) \cdot b(t_1, n_1, p_u) \cdot b(t_2, n_2, p_u),$$

where

$$f_l(t_1, t_2) = \begin{cases} 0, & \text{if } t_1 + t_2 \le r_2 \text{ or } t_1 \le r_1, \\ 1, & \text{if } t_1 + t_2 > s_2 \text{ or } t_1 > s_1, \\ B(r_3^* - t_1 - t_2, n_3^*, p_l), & \text{otherwise.} \end{cases}$$

We solve for r_3^* and s_3^* with the same approach as the two-stage cases.

By applying the conditional error function technique, we re-calculate new decision boundaries at final stage for two-stage designs and three-stage designs when the sample size is only changed at the final analysis. As a result, our proposed design allows for a completely free sample size modification at every stage while controlling the overall type I error rate. The examples are given in Section 5.

4. Software

We provide an R package that calculates the proposed two-/three-stage designs. This section gives a general sense of this package and introduces some basic usage. The complete document is in CRAN R documentation (Guo et al., 2019). To install `tsdf`, run the following command in R console:

```
> install.packages("tsdf")
```

The main function performing Phase II designs is `opt.design`. We will briefly go over this function, see some basic operations and have a look at the outputs. `opt.design` requires at least five inputs: `alpha1` as the left-side type I error, `alpha2` as the right-side type I error, `beta` as the type II error, `pc` as the minimal effective rate used in null hypothesis which can be a single value or an interval, and `pe` as the expected response rate. So a simple example would be

```
> opt.design(alpha1 = 0.15, alpha2 = 0.10, beta = 0.15, pc = 0.25,
 pe = 0.45)
```

The above code returns an object that contains all feasible designs and prints out the optimal one as below.

```
Minimal response rate: 0.25
Postulate response rate: 0.45
Left-side type 1 error: 0.15
Right-side type 1 error: 0.1
Type 2 error: 0.15
Notation:
Left-side rejection region at stage i is response <= ri
Right-side rejection region at stage i is response > si
Sample size used at stage i is ni
Optimal design :
r1 r2 s1 s2 n1 n2
 0 3 7 11 7 26
True errors :
  alpha11   alpha12   alpha21   alpha22     beta
0.13348389 0.14332822 0.00000000 0.09665671 0.12794521
```

By default, this function calculates two-stage designs that do not include early stop for superiority and do not apply alpha-spending method. Other key options include

- `stage`: A single value indicates whether two or three stage designs should be returned.
- `stop.eff`: A logical flag indicates if this trial can allow early stopping for efficacy.
- `sf.param`: A single real value specifying the gamma parameter for which Hwang-Shih-DeCani spending is to be computed.

`tsdf` supports Hwang-Shih-DeCani (Hwang et al., 1990) $\alpha-$spending function, which takes the form:

$$f(t, \alpha, \gamma) = \alpha(1 - \exp(-t\gamma))/(1 - \exp(-\gamma)),$$

where α is the overall type I error, t is the values of the proportion of sample size/information for which the spending function will be computed, and γ is a parameter that controls how the α is distributed at each stage. In function `dec.table`, `sf.param` specifies the choice of γ. Increasing γ implies that more error is spent at early stage and less is available in late stage. For example, a value of $\gamma = -4$ is used to approximate an O'Brien–Fleming design (O'Brien & Fleming, 1979), while a value of $\gamma = 1$ approximates a Pocock design (Jennison & Turnbull, 2000). We set the maximum sample size to be 100 by default as it may take more time to compute when n

Table 1. Differences between Simon's two-stage design and the new design.

	Simon's design	New design
Setup of threshold	Some uninteresting level, p_0. Some interesting level, p_1	The minimal effective response rate p_c or an interval $[p_l, p_u]$
Hypothesis setup	$H_0 : p \leq p_0$ vs. $H_1 : p \geq p_1$	$H_0 : p = p_c$ vs. $H_1 : p \neq p_c$ or $H_0 : p \in [p_l, p_u]$ vs. $H_1 : p \notin [p_l, p_u]$
Sample size at interim	Fixed or limited	Flexible
Trial conclusions	Inefficacy (no-go, $p < p_1$)	Inefficacy (no-go, $p < p_c$ or $p < p_l$) Efficacy (go, $p > p_c$ or $p > p_u$)

is large. It can be specified by the user based on their design settings and computing power. More details of this package can be found in the R documentation (Guo et al., 2019).

5. Evaluation

Simon's two-stage design (Simon, 1989) is the most commonly used design for Phase II oncology trials and any new method used in a trial may be asked by the question from investigators and IRBs (institutional review boards) to explain why the new design is needed and the difference between them. We summarize the differences in Table 1. In this section, we implement the proposed two-/three-stage designs and illustrate the procedure involved.

We look at two case studies, one with minimal response rate as single point and one as an interval.

Case 1: The minimal effective response rate is 0.40. The postulate response rate is at least 0.55. The type I error to conclude the compound is ineffective (conclude $p < 0.4$) when $p = 0.4$ and effective (conclude $p > 0.4$) when $p = 0.4$ are $0.3, 0.1$, respectively. The power to conclude the compound is effective (conclude $p > 0.4$) when $p = 0.55$ is at least 0.80.

We used Pocock spending function and calculated designs with and without early stopping for efficacy. We increase the total sample size n by 1 from 2 and stop until we find there are at least five choices of n_1 in the range of 30–60% of the total.

Table 2(a) summarizes two-stage designs without early stopping for efficacy. The study will enroll a total of 50 response-evaluable subjects, with stage 1 sample size between 15 and 30. The response-evaluable subjects will consist of all subjects who receive at least 1 dose of study drug and have at least one post-treatment disease assessment. At the end of the second stage, efficacy is declared when there are more than 24 responders from 50 response-evaluable subjects while inefficacy is declared when there are less than or equal to 17 responders. In this example, the boundaries at the second stage are not impacted by the stage 1 sample size. In fact, r_2, s_2 usually do not change much as the total sample size remains the same when the stage 1 sample size varies. The optimal design is when the stage 1 sample size is 22, with the expected sample size of 45.564. Table 3 provides designs when sample size at stage 2 may be changed from $n_2 = 26$ to $n_2^* = 25 \sim 31$ while the stage 1 sample size remains the same. As can be observed, the new boundaries r_2^* and s_2^* are adjusted along with the new sample size n_2^*. Namely, efficacy and inefficacy are claimed with smaller number of responders in the under-running cases while larger number of responders are needed to reach conclusion in the over-running cases. Table 2(b) presents two-stage designs with early stopping for efficacy. The total sample size is also 50. The stage 1 sample size has a range of 15–30. The optimal design is the same except adding efficacy boundary at stage 1. For example, when stage 1 sample size $n_1 = 15$, efficacy is declared when there are more than 11 responders. The inefficacy boundaries are similar to the designs without early stopping for efficacy in Table 2.

Case 2: The minimal effective response rate is between 0.40 and 0.45. The expected response rate is at least 0.60. The type I errors to conclude the compound is ineffective (conclude $p < 0.4$) when $p = 0.4$ and effective (conclude $p > 0.45$) when $p = 0.45$ are $0.3, 0.1$, respectively. The power to conclude the compound is effective(conclude $p > 0.4$) when $p = 0.60$ is at least 0.80.

Using the assumptions in Case 2, the study will need 53 response-evaluable subjects, with stage 1 sample size between 15 and 32. The inefficacy boundary for stage 1 varies from 3 to 10. At the end of the second stage, efficacy is declared when there are more than 28 responders from 53 response-evaluable subjects and inefficacy is declared when there are less than or equal to 18 responders. The optimal design is when the stage 1 sample size is 22 with an expected sample size of 48.088. The maximum expected sample size is 50.633 when stage 1 sample size is 19. The improvement in the expected sample size is very small while there is a big difference in stage 1 sample size. This indicates it is not necessary to seek the optimal design.

Table 2. Two-stage designs for Case 1: the minimal effective response rate is 0.40. The postulate response rate is at least 0.55. The type I errors to conclude the compound are ineffective (conclude $p < 0.4$) when $p = 0.4$ and effective (conclude $p > 0.4$) when $p = 0.4$ are 0.3, 0.1, respectively. The power to conclude the compound is effective (conclude $p > 0.4$) when $p = 0.55$ is at least 0.80.

(a) No early stopping for efficacy.

α_{11}	α_{12}	α_{22}	β	r_1	r_2	s_2	n_1	n_2	$E(n)$
0.090	0.272	0.097	0.198	3	17	24	15	35	46.832
0.065	0.259	0.098	0.197	3	17	24	16	34	47.785
0.126	0.286	0.097	0.198	4	17	24	17	33	45.842
0.094	0.269	0.098	0.197	4	17	24	18	32	46.987
0.070	0.257	0.098	0.197	4	17	24	19	31	47.842
0.126	0.281	0.098	0.197	5	17	24	20	30	46.232
0.096	0.266	0.098	0.197	5	17	24	21	29	47.224
0.158	**0.294**	**0.098**	**0.197**	**6**	**17**	**24**	**22**	**28**	**45.564**
0.124	0.275	0.098	0.197	6	17	24	23	27	46.653
0.096	0.262	0.098	0.197	6	17	24	24	26	47.505
0.154	0.285	0.098	0.197	7	17	24	25	25	46.161
0.122	0.269	0.098	0.197	7	17	24	26	24	47.083
0.184	0.297	0.098	0.197	8	17	24	27	23	45.769
0.148	0.278	0.098	0.197	8	17	24	28	22	46.733
0.119	0.263	0.098	0.197	8	17	24	29	21	47.507
0.176	0.287	0.098	0.197	9	17	24	30	20	46.474

(b) With early stopping for efficacy.

α_{11}	α_{12}	α_{21}	α_{22}	β	r_1	r_2	s_1	s_2	n_1	n_2	$E(n)$
0.090	0.272	0.002	0.098	0.198	3	17	11	24	15	35	46.765
0.065	0.259	0.005	0.099	0.196	3	17	11	24	16	34	47.619
0.126	0.286	0.000	0.097	0.198	4	17	13	24	17	33	45.827
0.094	0.269	0.006	0.099	0.196	4	17	12	24	18	32	46.803
0.070	0.257	0.003	0.098	0.196	4	17	13	24	19	31	47.747
0.126	0.281	0.006	0.099	0.196	5	17	13	24	20	30	46.038
0.096	0.266	0.004	0.098	0.196	5	17	14	24	21	29	47.120
0.158	**0.294**	**0.007**	**0.099**	**0.196**	**6**	**17**	**14**	**24**	**22**	**28**	**45.366**
0.124	0.275	0.004	0.098	0.196	6	17	15	24	23	27	46.546
0.096	0.262	0.007	0.099	0.196	6	17	15	24	24	26	47.310
0.154	0.285	0.004	0.098	0.197	7	17	16	24	25	25	46.053
0.122	0.269	0.008	0.099	0.196	7	17	16	24	26	24	46.894
0.184	0.297	0.005	0.098	0.197	8	17	17	24	27	23	45.663
0.148	0.278	0.008	0.099	0.196	8	17	17	24	28	22	46.555
0.119	0.263	0.013	0.100	0.195	8	17	17	24	29	21	47.224
0.176	0.287	0.008	0.099	0.196	9	17	18	24	30	20	46.308

Table 3. Two-stage designs for Case 1 with over-running and under-running at stage 2 (no stopping for efficacy). The planned design has $n_1 = 22, n_2 = 28$. The sample size at stage 2 is changed from 28 to 25–27 (under-running) and 29–31 (over-running).

α_{11}	α_{12}	α_{22}	β	r_1	r_2^*	s_2^*	n_1	n_2^*
Under-running at stage 2								
0.158	0.238	0.082	0.245	6	15	23	22	25
0.158	0.275	0.060	0.290	6	16	24	22	26
0.158	0.254	0.077	0.241	6	16	24	22	27
Planned design								
0.158	0.294	0.098	0.197	6	17	24	22	28
Over-running at stage 2								
0.158	0.271	0.073	0.236	6	17	25	22	29
0.158	0.250	0.092	0.195	6	17	25	22	30
0.158	0.288	0.070	0.232	6	18	26	22	31

6. Summary and discussion

This paper presented modifications and extensions of the two-stage designs of B. Zhong (2012) to solve some practical issues when conducting single-arm Phase II oncology trials. We provide flexible two-/three-stage designs which allows flexible interim sample size, early stopping rule for efficacy, hypothesized interval of response rate when a single minimal response rate is not available, and handling of over-running and under-running while controlling overall type I error. Moreover, we provide an open-source R package that integrates the aforementioned features.

In Zhong's design, an inconclusive outcome is added as a natural outcome of the designs. The inconclusive outcome corresponds to a trial where definitive go/no-go decisions cannot be made. This inconclusive result is an unavoidable result due to the small sample size as well as the intrinsic nature of these uncontrolled clinical trials.

Table 4. Two-stage designs for Case 2: the minimal effective response rate is between 0.40 and 0.45. The expected response rate is at least 0.60. The type I errors to conclude the compound is ineffective (conclude $p < 0.4$) when $p = 0.4$ and effective (conclude $p > 0.45$) when $p = 0.45$ are 0.3, 0.1, respectively. The power to conclude the compound is effective (conclude $p > 0.4$) when $p = 0.60$ is at least 0.80.

α_{11}	α_{12}	α_{22}	β	r_1	r_2	s_2	n_1	n_2	$E(n)$
0.090	0.263	0.100	0.177	3	18	28	15	38	49.561
0.065	0.250	0.100	0.177	3	18	28	16	37	50.590
0.126	0.279	0.100	0.177	4	18	28	17	36	48.464
0.094	0.261	0.100	0.177	4	18	28	18	35	49.704
0.070	0.248	0.100	0.177	4	18	28	19	34	50.633
0.126	0.274	0.100	0.177	5	18	28	20	33	48.855
0.096	0.258	0.100	0.177	5	18	28	21	32	49.936
0.158	**0.288**	**0.100**	**0.177**	**6**	**18**	**28**	**22**	**31**	**48.088**
0.124	0.268	0.100	0.177	6	18	28	23	30	49.281
0.096	0.254	0.100	0.177	6	18	28	24	29	50.217
0.154	0.280	0.100	0.177	7	18	28	25	28	48.701
0.122	0.263	0.100	0.177	7	18	28	26	27	49.718
0.184	0.292	0.100	0.177	8	18	28	27	26	48.217
0.148	0.272	0.100	0.177	8	18	28	28	25	49.288
0.143	0.264	0.100	0.177	9	18	28	31	22	49.846
0.205	0.293	0.100	0.177	10	18	28	32	21	48.704

Our design can be further extended by integrating an additional hypothesis testing with the secondary endpoint and dividing the ‘grey zone’ into a ‘sub-go zone’ and a ‘sub-no-go zone’. Moreover, the inefficacy/efficacy boundaries are established with the new two-stage design. The concept of inefficacy is used instead of the commonly used term ‘futility’. The term ‘futility’ in clinical trials is used to refer to the inability of a clinical trial to achieve its objectives. For example, a trial is designed with 65% response rate for the test treatment group and 50% response rate for the control group (standard of care) and with the intention to establish the superiority. Therefore, the study objective is to demonstrate that the test treatment is superior over the standard of care. In the middle of the study, an interim analysis was performed and revealed 55% response rate for the new treatment group and 50% response rate for the control group. Based on the results from the interim analysis, it is concluded that it is unlikely to reach statistical significance at the end of the study. Therefore, the trial is terminated. Note that the test treatment is still better than the standard of care by the interim observed response rates. If the sample size is large enough, statistical significance may still be demonstrated. In other words, futility does not mean inefficaciousness. For this example, the observed response rate at the interim analysis is 55% and this is higher than the 50% response rate of the standard of care. The test treatment is still effective even if it has a 50% response rate. It is obvious that ineffectiveness cannot be concluded from the interim data but futility conclusion can be made.

The designs presented in this paper can also be used for multi-arm trials with the intention to identify/rule out ineffective doses/treatments where go/no-go decisions are based on the data from individual arms. Finally, the framework of the paper is based on hypothesis testing. Other extensions such as using confidence intervals approach or extending our framework to other type of endpoints and randomized trials will be investigated in the future.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

Brannath, W., Koenig, F., & Bauer, P. (2007). Multiplicity and flexibility in clinical trials. *Pharmaceutical Statistics*, *6*(3), 205–216. https://doi.org/10.1002/pst.302

Demets, D. L., & Lan, K. K. G. (1994). Interim analysis: The alpha spending function approach. *Statistics in Medicine*, *13*(13–14), 1341–1352. https://doi.org/10.1002/(ISSN)1097-0258

Englert, S., & Kieser, M. (2012). Improving the flexibility and efficiency of phase II designs for oncology trials. *Biometrics*, *68*(3), 886–892. https://doi.org/10.1111/j.1541-0420.2011.01720.x

Englert, S., & Kieser, M. (2015). Methods for proper handling of overrunning and underrunning in phase II designs for oncology trials. *Statistics in Medicine*, *34*(13), 2128–2137. https://doi.org/10.1002/sim.v34.13

Guo, W., Hui, J., & Zhong, B. (2019). *TSDF: Two-/three-stage designs for phase 1&2 clinical trials*. R package version 1.1-7.

Hong, S., & 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/(ISSN)1097-0258

Hwang, I. K., Shih, W. J., & De Cani, J. S. (1990). Group sequential designs using a family of type i error probability spending functions. *Statistics in Medicine*, *9*(12), 1439–1445. https://doi.org/10.1002/(ISSN)1097-0258

Jennison, C., & Turnbull, B. W. (2000). *Group sequential methods with applications to clinical trials*, Chapman and Hall.

Koyama, T., & Chen, H. (2008). Proper inference from Simon's two-stage designs. *Statistics in Medicine*, *27*(16), 3145–3154. https://doi.org/10.1002/sim.v27:16

Li, G., Shih, W. J., Xie, T., & Lu, J. (2002). A sample size adjustment procedure for clinical trials based on conditional power. *Biostatistics*, *3*(2), 277–287. https://doi.org/10.1093/biostatistics/3.2.277

O'Brien, P. C., & Fleming, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics*, *35*(3), 549–556. https://doi.org/10.2307/2530245

Proschan, M. A., & Hunsberger, S. A. (1995). Designed extension of studies based on conditional power. *Biometrics*, *51*(4), 1315–1324. https://doi.org/10.2307/2533262

Sargent, D. J., Chan, V., & 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

Shan, G., & Chen, J. J. (2018). Optimal inference for Simon's two-stage design with over or under enrollment at the second stage. *Communications in Statistics – Simulation and Computation*, *47*(4), 1157–1167. https://doi.org/10.1080/03610918.2017.1307398

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

Whitehead, J. (1992). Overrunning and underrunning in sequential clinical trials. *Controlled Clinical Trials*, *13*(2), 106–121. https://doi.org/10.1016/0197-2456(92)90017-T

Zhong, B. (2012). Single-arm phase IIa clinical trials with go/no-go decisions. *Contemporary Clinical Trials*, *33*(6), 1272–1279. https://doi.org/10.1016/j.cct.2012.07.006

Zhong, W., & Zhong, B. (2013). One-sample proportion testing procedure for hypothesis of inequality. *Journal of Biopharmaceutical Statistics*, *23*(3), 604–617. https://doi.org/10.1080/10543406.2012.756501