Received: 9 June 2020 | Revised: 21 September 2020 | Accepted: 26 November 2020

DOI: 10.1002/pst.2089

MAIN PAPER

WILEY

A confidence function-based posterior probability design for phase II cancer trials

Minghua Shan

Bayer U.S. LLC Pharmaceuticals, Whippany, New Jersey, USA

Correspondence
Minghua Shan, Bayer U.S. LLC, Pharmaceuticals, 100 Bayer Boulevard, Whippany, NJ 07981.
Email: minghua.shan@bayer.com

Abstract

Single-arm one- or multi-stage study designs are commonly used in phase II oncology development when the primary outcome of interest is tumor response, a binary variable. Both two- and three-outcome designs are available. Simon two-stage design is a well-known example of two-outcome designs. The objective of a two-outcome trial is to reject either the null hypothesis that the objective response rate (ORR) is less than or equal to a pre-specified low uninteresting rate or to reject the alternative hypothesis that the ORR is greater than or equal to some target rate. Three-outcome designs proposed by Sargent et al. allow a middle gray decision zone which rejects neither hypothesis in order to reduce the required study size. We propose new two- and three-outcome designs with continual monitoring based on Bayesian posterior probability that meet frequentist specifications such as type I and II error rates. Futility and/or efficacy boundaries are based on confidence functions, which can require higher levels of evidence for early versus late stopping and have clear and intuitive interpretations. We search in a class of such procedures for optimal designs that minimize a given loss function such as average sample size under the null hypothesis. We present several examples and compare our design with other procedures in the literature and show that our design has good operating characteristics.

KEYWORDS

Bayesian posterior probability, continual monitoring, phase II clinical trial, sample size, two-stage design

1 | INTRODUCTION

In oncology drug development, phase II studies are typically conducted based on a surrogate endpoint to screen for potentially efficacious treatments. Single-arm one- or multi-stage study designs are commonly used when the primary outcome of interest is tumor response, a binary variable. Simon two-stage procedures[1] are often used in this setting. These procedures allow early stopping after stage 1 for futility when the experiment treatment is unlikely to have the targeted response rate and therefore reduce the number of patients treated with a toxic but ineffective therapy. Sargent et al. proposed one- and two-stage designs[2] that have three possible outcomes in order to reduce the number of patients needed. In addition to the outcomes of rejecting the null or alternative hypothesis, these three-outcome procedures

allow a middle gray zone which rejects neither. If a study outcome falls in this inconclusive decision zone, one may need to consider additional information in making a decision regarding whether and how to move forward in developing the experimental therapy. As many valid study designs meet any given error rate and power specifications, one may choose to use a procedure that is optimal in some sense such as minimizing the maximum sample size or minimizing the average sample size under the null hypothesis or some other meaningful loss function.

Three-stage designs with and without early efficacy boundaries have also been proposed in the literature.[3,4] In general, multi-stage designs are more efficient in the sense that they tend to have a smaller average sample size due to the possibility of early stopping. However, these multi-stage procedures are rigid as, for a given stage, the sample size and decision boundaries are fixed. In practice, the actual number of patients for a given stage is often not exactly the same as planned. If this happens, the decision boundary would not be clearly defined and the operating characteristics of any decision based on the study data would not be known. Green and Dahlberg[5] investigated several empirical approaches to adapting stopping rules in this situation. Koyama and Chen[6] proposed a method for inferences using both planned and actual sample sizes. Lee and Liu[7] proposed a two-outcome sequential design based on Bayesian predictive probability (PP) that allows continual monitoring after each additional patient and meets pre-specified error rate requirements. The design can have stopping boundaries for futility and/or efficacy and can potentially reject the null or the alternative hypothesis at any time during the study. The authors compared these more flexible procedures with the Simon two-stage design when either procedures have early stopping for efficacy and concluded that the predictive probability design may have a smaller average sample size under the null hypothesis.

In this paper, we propose a new design that meets typical frequentist specifications, such as type I and II error rates, and allows continual monitoring. It may have early boundaries for futility and/or efficacy. Boundaries are based on Bayesian posterior probabilities and confidence functions (CF). It has the flexibility to require varying levels of evidence for early stopping (e.g., requiring a higher level of confidence that the therapy is efficacious in order to stop early compared to reaching such conclusion at the final analysis). An optimal design can be obtained by searching a group of valid procedures to minimize a given loss function such as average sample size under the null hypothesis.

In Section 2, we formally define our three-outcome posterior probability design based on confidence functions. We also outline a search algorithm for finding such procedures. Section 3 presents a two-outcome design as a simple derivative of our three-outcome design. We search for an optimal design among a class of valid procedures that minimizes a given loss function. Finally, in Section 4, we provide several examples and compare our design with Simon, Sargent et al., and the predictive probability design by Lee et al.

2 | THREE-OUTCOME PROCEDURES

In the phase II setting of oncology drug development, let $p \in (0, 1)$ be the underlying objective response rate (ORR) of interest for an experimental therapy. We design a single-arm trial to test

$$H_0 : p \leq p_0$$

versus

$$H_a : p \geq p_a$$

where p_0 is the response rate of standard of care and p_a is the target response rate for the experimental treatment. Following Sargent et al.,[2] we define

$$\alpha = \max_{p \in (0, p_0]} Pr(\text{reject} H_0 | H_0 \text{is true})$$
$$\beta = \max_{p \in [p_a, 1)} Pr(\text{reject} H_a | H_a \text{is true})$$
$$\eta = \min_{p \in (0, p_0]} Pr(\text{reject} H_a | \text{H}_0 \text{is true})$$
$$\pi = \min_{p \in [p_a, 1)} Pr(\text{reject H}_0 | \text{H}_a \text{is true})$$

That is, α and β are the false positive and false negative error rates, π is the probability of correctly rejecting H_0 (i.e., power), and η is the probability of correctly rejecting H_a (when H_0 is true). Note that $1-\alpha-\eta$ and $1-\beta-\pi$ represent the probabilities of landing in the inconclusive or gray decision zone under H_0 and H_a, respectively.

Let n be the maximum planned sample size for a study. Let $\tilde{X}_1, \tilde{X}_2, ..., \tilde{X}_n$ represent responses for the first, second, ..., and the n^{th} patient with $\tilde{X}_i = 1$ if the i^{th} patient is a responder and 0 if non-responder. Define $X_i = \sum_{j=1}^{i} \tilde{X}_j$, the number of responders up to patient i. For any given $i = 1, 2, ..., n$, X_i follows a binomial distribution with i trials and probability of success p, $Bin(i, p)$. We let r_i and s_i ($r_i < s_i$) be integers representing boundaries for rejecting H_a and H_0, respectively, after patient i, $i = 1, 2, ..., n$. That is, we reject H_a if $X_i \le r_i$ and reject H_0 if $X_i \ge s_i$. Define

$$p_i^L(p) = Pr\left(X_i \le r_i, r_j < X_j < s_j \text{ for } 1 \le j \le i-1 \mid p\right)$$
$$p_i^U(p) = Pr\left(X_i \ge s_i, r_j < X_j < s_j \text{ for } 1 \le j \le i-1 \mid p\right)$$

Then p_i^L is the probability of crossing the futility boundary at the i^{th} patient when the true response rate is p and that neither futility nor efficacy boundary was crossed before. Similarly, p_i^U is the corresponding probability of crossing the efficacy boundary at the i^{th} patient.

It is immediately clear that

$$\begin{aligned} \alpha &= \sum_{i=1}^{n} p_i^U(p_0) \\ \beta &= \sum_{i=1}^{n} p_i^L(p_a) \\ \eta &= \sum_{i=1}^{n} p_i^L(p_0) \\ \pi &= \sum_{i=1}^{n} p_i^U(p_a) \end{aligned} \tag{1}$$

When designing a trial, we wish to find appropriate values of n, r_i, and s_i ($1 \le i \le n$) so that α and β are less than or equal to their respective pre-specified values and η and π are greater than or equal to theirs. Any procedure specified by $\{n, r_i, s_i, i = 1, 2, ..., n\}$ that meets these specifications is a valid study design.

2.1 | Bayesian posterior probability and stopping boundaries

Suppose p follows a Beta distribution. We assume a non-informative prior of $Beta(0.5,0.5)$. However, when appropriate one could consider using a more informative prior based on data already available. At the time the i^{th} patient has been treated and followed sufficiently for response assessment, the posterior distribution of p is $Beta(0.5 + X_i, 0.5 + i - X_i)$. Let $b(\theta \mid X_i, i)$ be the corresponding density function. The posterior probability that response rate p is lower than the target p_a, or H_a is not true, is

$$PP_a(X_i, i) = Pr(p < p_a | X_i, i) = \int_0^{p_a} b(\theta | X_i, i) d\theta \tag{2}$$

Similarly, the posterior probability that response rate p is greater than the null value of p_0, or H_0 is not true, is

$$PP_0(X_i, i) = Pr(p > p_0 | X_i, i) = \int_{p_0}^{1} b(\theta | X_i, i) d\theta \tag{3}$$

We formulate stopping boundaries r_i and s_i based on $PP_a(X_i, i)$ and $PP_0(X_i, i)$ respectively. Specifically, let ω_i^U and $\omega_i^L \in (0,1]$ for $i = 1, 2, ..., n$. We define s_i so that

$$s_i = \underset{0 \le x \le i}{min} \left\{ PP_0(x,i) \ge \omega_i^U \right\},\ i = 1,2,...,n \tag{4}$$

In other words, s_i is the minimum number of responses needed in the first i patients such that the posterior probability that H_0 is not true is $\ge \omega_i^U$. If no value of x exists that satisfies Equation (4), we let $s_i = i + 1$, which means no stopping for efficacy at i patients as X_i is never greater than the number of patients. Note that if $\omega_i^U = 1$, we will not be able to stop for efficacy at i.

Analogously, we define

$$r_i = \underset{0 \le x \le i}{max} \left\{ PP_a(x,i) \ge \omega_i^L \right\},\ i = 1,2,...,n \tag{5}$$

If no x satisfies Equation (5), we let $r_i = -1$, meaning no stopping for futility at i. This way, stopping boundaries $\{r_i, s_i, i = 1, 2, ..., n\}$ are completed determined after we specify ω_i^U and ω_i^L, $i = 1, 2, ..., n$. Note that ω_i^U and ω_i^L are the minimum threshold posterior probabilities, or confidence levels, for rejecting H_0 and H_a at patient i.

2.2 | Confidence functions

We introduce confidence functions (CF) for determining futility and/or efficacy boundaries described above. Let $f(t, c)$ be a function defined for $t \in (0, 1]$ and $c \in (0, 1]$ that satisfies the following conditions:

1. $0 < f(t, c) \le 1$ for all t and c
2. For any give value of t, $f(t, c)$ is continuous and monotone non-decreasing in c
3. For any give value of t, $\lim_{c \to 1} f(t, c) = 1$
4. $f(t = 1, c)$ is strictly increasing in c

We shall call such functions confidence functions. Consider a function in the above family. When c is given, we can visualize a curve plotting $f(t, c)$ against t over the interval (0, 1]. As c increases, the curve is lifted upward (even though its shape might change), and as c approaches 1, the curve gets closer and closer to the horizontal straight line at 1.

It is easily seen that the following functions meet the above conditions.

$$f(t,c) = c, \qquad t \in (0,1] \text{ and } c \in (0,1] \tag{6}$$

$$\begin{aligned} f(t,c) &= 1 - t^3(1-c), && t \in (0,1] \text{ and } c \in (0,1] \\ f(t,c) &= I(t<1) + cI(t=1), && t \in (0,1] \text{ and } c \in (0,1] \end{aligned} \tag{7}$$

where $I(.)$ is an indicator function.

Such confidence functions can also be formed based on many common α-spending functions as shown below:

$$f(t,c) = 1 - g(t, 1-c)$$

where $g(t, \alpha)$ is an α-spending function. A useful family of confidence functions is

$$f_\gamma(t,c) = \begin{cases} 1 - \left[\dfrac{1 - exp(-\gamma t)}{1 - exp(-\gamma)} \right](1-c), & \text{if } \gamma \ne 0 \\ 1 - t(1-c), & \text{if } \gamma = 0 \end{cases} \tag{8}$$

for $t \in (0, 1]$ and $c \in (0, 1]$. Equation (8) is based on the gamma-family error spending functions introduced by Hwang et al.[8] Note that Equation (8) is a strictly decreasing function in t for any given c and γ. The decrease is steeper for t close to 1 when γ has a large negative value. On the other hand, when γ has a large positive value, the decrease is more rapid

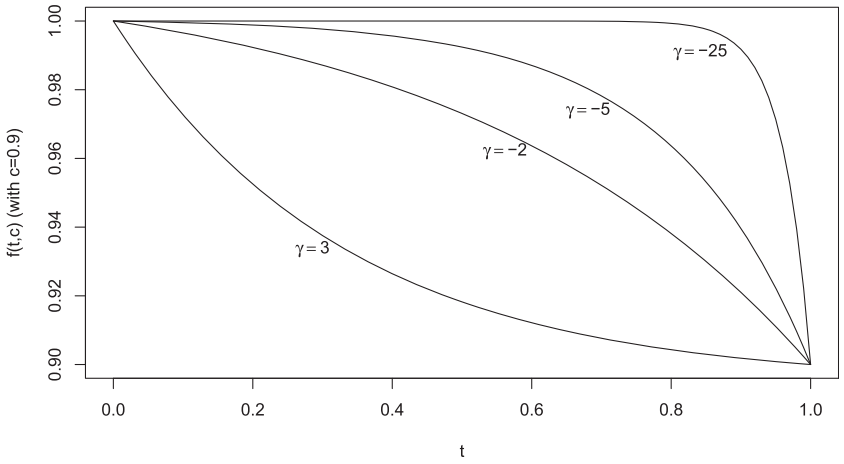

FIGURE 1 Plot of confidence functions in Equation (8) for select values of γ and $c = 0.9$

when t is close to 0. See Figure 1. The family of functions in Equation (8) is useful because it can represent many diversified patterns. As can be seen later in Section 2.3, this allows for study boundaries corresponding to varying levels of evidence before early stopping, and c is the threshold posterior probability for the final analysis. We note that functions in Equation (8) converge to Equation (6) when $\gamma \to \infty$, and converge to Equation (7) when $\gamma \to -\infty$. It will become clear that the latter is useful in constructing our study design when no early stopping for efficacy (or futility) is desired.

2.3 | Boundaries based on confidence functions

Let n be the maximum number of patients planned for a study. Given confidence functions $f^U(t, c)$ and $f^L(t, c)$ for efficacy and futility boundaries respectively, for any values c^U and c^L, we let

$$\begin{aligned} \omega_i^U &= f^U\left(i/n, c^U\right),\ i = 1, 2, ..., n \\ \omega_i^L &= f^L\left(i/n, c^L\right),\ i = 1, 2, ..., n \end{aligned} \tag{9}$$

Here i/n can be viewed as the information time, and confidence functions resemble error-spending functions for a group sequential design but have very different interpretations. From ω_i^L and ω_i^U, we obtain stopping boundaries r_i and s_i based on Equations (5) and (4) in Section 2.1. Note that r_i and s_i are functions of c^L and c^U as well as n. For a given value of n, r_i and s_i go up or down when c^L and c^U increase or decrease. When s_i decreases, the false positive rate α increases. Similarly, the false negative rate β increases as r_i increases. Therefore, we can adjust c^L and c^U to obtain r_i and s_i that meet α and β specifications. Additionally, we can increase the total study size n so that the η and π requirements are also met.

When confidence function (6) or (8) with a large positive γ is used, the resulting threshold posterior probability for stopping is constant with respect to the timing of analysis. However, when conducting a clinical study, we typically would consider early stopping only if early evidence is overwhelmingly convincing. Therefore, it is generally recommendable to design trials so that early stopping boundaries require a higher level of confidence than late boundaries. For our CF design, this can be achieved by using a confidence function that is decreasing in t. Several functions described in Section 2.2, including Equation (8), possess this property.

For confidence functions in Equation (8), we note that $f(1, c) = c$. Therefore the parameter c is the threshold posterior probability, ω_n^U or ω_n^L, required for rejecting H_0 or H_a at the final analysis with n patients. This shows that parameter c in our confidence function $f(t, c)$ in Equation (8) has a clear and intuitive interpretation.

2.4 | Search algorithm

For given confidence functions $f^U(t, c)$ and $f^L(t, c)$, we search for the minimum value of n such that stopping boundaries r_i and s_i as defined in Section 2.3 meet all specifications regarding α, β, η, and π.

Step 0: Let $n = 1$.

Step 1: Find values of c for $f^U(t, c)$ and $f^L(t, c)$ respectively so that the corresponding efficacy and futility boundaries, as defined in Equations (4), (5), and (9), meet α and β specifications.

1a: Assuming no futility boundary (i.e., $r_i = -1$, $i = 1, 2, ..., n$), find the value of $c = c^U$ such that the corresponding efficacy boundary $\{s_i, i = 1, 2, ..., n\}$ meets the α requirement (i.e., less than or equal to the specified α value). Because $f^U(t, c)$ is a monotone function in c, for a given efficacy boundary corresponding to c, its α value is a monotone function of c. Therefore we can perform a binary search for c^U.

1b: With the above efficacy boundary in place, find the value of $c = c^L$ such that the corresponding futility boundary $\{r_i, i = 1, 2, ..., n\}$ meets the β requirement by performing a binary search similar to **1a**.

1c: Given the new futility boundary, the actual α corresponding to the current efficacy boundary may be reduced and become smaller than necessary. With the new futility boundary, update c^U so that α is as large as possible but still meets specification.

1d: Because the new efficacy boundary corresponding to the updated c^U may be lower (therefore more likely to cross), actual β corresponding to the current futility boundary may be reduced (because we cannot cross the futility boundary if we already crossed the efficacy boundary earlier) and become smaller than necessary. With the new efficacy boundary, update c^L so that β is as large as possible but still meets specification.

1e: Repeat **1c** and **1d** a sufficient number of times so that c^U and c^L converge to stable values and note the corresponding boundaries $\{s_i, i = 1, 2, ..., n\}$ and $\{r_i, i = 1, 2, ..., n\}$.

Step 2: Check whether the boundaries from **Step 1** meet the power specifications regarding η and π.

2a: For the boundaries from **Step 1**, obtain η and π as shown in Equation (1).

2b: If either η or π does not meet specification (i.e., less than the desired value), increase n by 1 and go to **Step 1**.

2c: If both η and π meet specifications, stop and we have found the minimum maximum study size n (for the given confidence functions) and futility and efficacy boundaries $\{r_i, s_i, i = 1, 2, ..., n\}$ that meet all error rate and power specifications.

2.5 | Procedures without early stopping for efficacy

Sometimes we may prefer a study design that only allows early stopping for futility without the possibility of early stopping for efficacy. We can achieve this by using confidence function (7) for $f^U(t, c)$ when searching for a procedure as described in Section 2.4. As the posterior probability required for early stopping for efficacy is 1 prior to the final analysis with n patients, there is no possibility of efficacy early stopping. Alternatively, we can use confidence function (8) with a large negative value for γ to achieve the same as Equation (8) converges to Equation (7) when $\gamma \to -\infty$.

2.6 | Group sequential procedures

The study design we described above is a sequential procedure which allows continual analysis of the accumulating study data after every additional patient. If for some reason, such as operational, we would like to start analyzing the data only after a pre-specified minimum number of patients and/or perform subsequent analyses only after a pre-specified number of additional patients have been included, we can use a group sequential design. The same search algorithm with a minor adjustment can be used to find a group sequential design. As in Equation (9), we let $\omega_i^U = f^U(i/n, c^U)$ and $\omega_i^L = f^L(i/n, c^L)$, but only for i values at which we plan to perform analyses; for other i values, we let $\omega_i^U = i + 1$ and $\omega_i^L = -1$ so that early stopping is made impossible.

3 | TWO-OUTCOME AND OPTIMAL PROCEDURES

3.1 | Two-outcome procedures

In Section 2, we presented a three-outcome procedure that includes an inconclusive decision zone in order to reduce the study size. The same method can be used to obtain a two-outcome design by requiring the following:

$$\eta = 1-\alpha$$
$$\pi = 1-\beta$$

Because this forces the probability of landing in the gray decision zone to be zero, the resulting procedure will only have two possible outcomes: rejecting H_0 or rejecting H_a.

3.2 | Optimal designs

As described above, for any given confidence function, a valid two- or three-outcome design with the smallest n possible can be obtained. When many confidence functions are acceptable for use in a given setting, we can search among the corresponding study designs for an optimal procedure that minimizes a given loss function. For example, we consider designs based on the family of confidence functions in Equation (8) indexed by the parameter γ. When multiple values of γ are considered reasonable, we can search for γ such that the corresponding design minimizes the average sample size under H_0, $E(N|\, p_0)$. Alternatively, one may choose to minimize the maximum study size (i.e., minimax). However, as one-stage designs typically have the smallest maximum sample sizes and they are valid CF procedures corresponding to very large negative values of γ, when searching for a minimax design, it is recommended to exclude very large negative values of γ in the search unless one is interested in a one-stage procedure. Other alternative loss functions can also be considered. Jung et al.[9] considered a loss function that is a weighted average of the maximum study size n and $E(N|\, p_0)$ that results in Bayesian admissible designs.

4 | EXAMPLES

As examples, we present several designs with and without an early efficacy boundary. We compare our confidence function (CF) based two-outcome procedures without early efficacy boundary with Lee et al.[7] predictive probability (PP) procedures and Simon[1] optimal and minimax designs. Three-outcome CF designs without early efficacy boundary are compared with Sargent et al.'s minimax and optimal designs. In this Section, "optimal" refers to a study design that minimizes the expected number of patients under H_0, $E(N|\, p_0)$. All optimal and minimax CF designs in this section were based on CF (8) and obtained using grid search over specified ranges of γ.

Example 1. *An optimal three-outcome procedure with early stopping boundaries for both efficacy and futility.*

We design a three-outcome single-arm clinical trial testing

$$H_0 : p \le p_0 = 0.2$$

versus

$$H_a : p \ge p_a = 0.4,$$

where p is the underlying true ORR of the experimental therapy. We require that $\alpha \le 0.1$, $\beta \le 0.1$, $\eta \ge 0.8$, and $\pi \ge 0.8$. The first analysis of the accumulating data is planned only after the tenth patient has been treated and followed for response. After that, data are analyzed continually after every additional patient. Suppose that, for the family of confidence functions (8), γ values between 0 and -10 are all considered acceptable for deriving both the futility and efficacy boundaries. Additionally, we allow the two boundaries to be based on different confidence functions and therefore different γ values, γ^L and γ^U. We search the two-dimensional grid defined by $-10 \le \gamma^L \le 0$ and $-10 \le \gamma^U \le 0$ with step size of 0.01 to find the design that minimizes $E(N|\, p_0)$, the average sample size under H_0.

The optimal CF procedure corresponds to $\gamma^L = -4.25$ and $\gamma^U = -0.95$. It has $E(N|\, p_0) = 17.39$. The maximum number of patients needed is 26. The minimum posterior probability as defined in Equation (2) for rejecting H_a at the final analysis with 26 patients is $\omega_n^L = c^L = 0.9158$. That required for rejecting H_0 at the final analysis is $\omega_n^U = c^U = 0.9457$. See Table 1 for more details. See Table 2 for efficacy and futility boundaries and probabilities of stopping at each analysis under H_0 and H_a, respectively.

Two-stage three-outcome procedures by Sargent et al. do not have an early efficacy boundary. For comparison, we obtain our procedure again without an early efficacy boundary by letting $\gamma^U = -\infty$. Our optimal procedure that minimizes $E(N|\ p_0)$ has $n = 27$ and $E(N|\ p_0) = 18.56$. Sargent et al.'s optimal and minimax designs have $E(N|\ p_0) = 20.97$ and 21.19, respectively. See Table 3 for more details, which also includes the CF minimax design. Table 4 shows boundaries and stopping probabilities for the CF optimal design without an early efficacy boundary.

Example 2. *An optimal two-outcome procedure without early stopping for efficacy.*

Lee et al.[7] devised a Bayesian predictive probability (PP) two-outcome procedure for testing the same hypotheses in Example 1 but without an early efficacy boundary and compared it with the Simon optimal and minimax two-stage designs. In addition to the flexibility of allowing continual monitoring of accumulating data, the authors found that the

α	β	η	π	γ^L	γ^U	n	c^L	c^U	$E(N\|\ p_0)$
0.0982	0.0976	0.8002	0.8106	−4.25	−0.95	26	0.9158	0.9457	17.39

TABLE 1 Optimal three-outcome CF design with both early futility and efficacy boundaries and its operating characteristics for Example 1

TABLE 2 Boundaries and stopping probabilities for optimal CF design with both early futility and efficacy boundaries in Example 1

i	r_i	s_i	ω_i^L	ω_i^U	Stopping probability under H_0	Stopping probability under H_a
10	0	5	0.9950	0.9849	0.1402	0.0388
12	1	6	0.9926	0.9812	0.1753	0.0180
13	1	6	0.9910	0.9792	0.0072	0.0072
15	2	7	0.9871	0.9750	0.1452	0.0158
16	2	7	0.9846	0.9728	0.0048	0.0048
17	2	7	0.9816	0.9705	0.0076	0.0076
18	3	7	0.9781	0.9682	0.1274	0.0231
20	4	8	0.9692	0.9631	0.1196	0.0215
21	4	8	0.9635	0.9605	0.0055	0.0055
22	4	8	0.9568	0.9578	0.0082	0.0082
23	5	9	0.9489	0.9549	0.0763	0.0138
24	5	9	0.9396	0.9520	0.0022	0.0022
25	6	9	0.9287	0.9489	0.0725	0.0228
26	6	9	0.9158	0.9457	0.1080	0.1166

TABLE 3 Comparison of CF and Sargent et al. three-outcome designs for Example 1

Design	α	β	η	π	γ^L	γ^U	n_1	n	c^L	c^U	$E(N\|\ p_0)$
CF optimal	0.0711	0.0989	0.8056	0.8062	−2.73	$-\infty$		27	0.9353	0.8911	18.56
CF minimax	0.0889	0.0999	0.8147	0.8074	−9.07	$-\infty$		24	0.9039	0.8677	19.80
Sargent optimal	0.0999	0.0931	0.8170	0.8560			13	29			20.97
Sargent minimax	0.0889	0.0987	0.8130	0.8070			16	24			21.19

expected number of patients for the predictive probability procedure ($E(N|\ p_0) = 27.67$) is smaller than that for the Simon minimax 2-stage design ($E(N|\ p_0) = 28.26$) but larger than the Simon "optimal" design ($E(N|\ p_0) = 26.02$). Lee et al. also included designs with larger maximum sample sizes, some of which have a lower $E(N|\ p_0)$. For example, the design with maximum $n = 42$ has $E(N|\ p_0) = 23.56$.

TABLE 4 Boundaries and stopping probabilities for optimal CF design without early efficacy boundary in Example 1

i	r_i	s_i	ω_i^L	ω_i^U	Stopping probability under H_0	Stopping probability under H_a
10	0	-	0.9921	1.0000	0.1074	0.0060
12	1	-	0.9893	1.0000	0.1718	0.0145
15	2	-	0.9839	1.0000	0.1429	0.0136
18	3	-	0.9767	1.0000	0.1168	0.0125
20	4	-	0.9704	1.0000	0.1169	0.0187
23	5	-	0.9583	1.0000	0.0777	0.0140
25	6	-	0.9480	1.0000	0.0721	0.0195
27	6	9	0.9353	0.8911	0.1944	0.9011

TABLE 5 Comparison of CF, PP, and Simon designs for Example 2

Design	α	β	γ^L	n	c^L	c^U	$E(N\mid p_0)$
CF optimal	0.097	0.100	−0.80	38	0.9581	0.8359	21.75
CF minimax	0.0847	0.0995	−5.88	36	0.9295	0.8763	24.84
Simon optimal	0.095	0.097		37			26.02
Simon minimax	0.086	0.098		36			28.26
PP with min n	0.088	0.094		36			27.67
PP with n = 42	0.099	0.083		42			23.56

TABLE 6 Boundaries and stopping probabilities for optimal CF design without early efficacy boundary for Example 2

i	r_i	s_i	ω_i^L	ω_i^U	Stopping probability under H_0	Stopping probability under H_a
10	0	-	0.9920	1.0000	0.1074	0.0060
12	1	-	0.9902	1.0000	0.1718	0.0145
15	2	-	0.9873	1.0000	0.1429	0.0136
19	3	-	0.9832	1.0000	0.0934	0.0075
22	4	-	0.9799	1.0000	0.0868	0.0078
25	5	-	0.9763	1.0000	0.0735	0.0075
28	6	-	0.9725	1.0000	0.0606	0.0069
30	7	-	0.9699	1.0000	0.0617	0.0106
33	8	-	0.9657	1.0000	0.0421	0.0081
36	9	-	0.9612	1.0000	0.0320	0.0069
38	10	11	0.9581	0.8359	0.1278	0.9105

TABLE 7 A group sequential CF design for Example 3

α	β	γ^L	γ^U	n	c^L	c^U	$E(N\mid p_0)$
0.0982	0.0967	4.60	$-\infty$	43	0.9712	0.8219	22.37

i	r_i	s_i	ω_i^L	ω_i^U	Stopping probability under H_0	Stopping probability under H_a
10	1	-	0.9809	1.0000	0.3758	0.0464
15	2	-	0.9768	1.0000	0.0990	0.0094
20	3	-	0.9744	1.0000	0.0621	0.0037
25	5	-	0.9729	1.0000	0.1460	0.0120
30	6	-	0.9721	1.0000	0.0390	0.0025
35	8	-	0.9716	1.0000	0.0858	0.0075
40	10	-	0.9713	1.0000	0.0692	0.0101
43	11	12	0.9712	0.8219	0.1231	0.9084

TABLE 8 Additional comparisons with two-stage two-outcome designs (Simon) for Example 4

	Two-stage two-outcome design (Simon) Optimal and Minimax					CF two-outcome design without early efficacy boundary Optimal and Minimax				
p_0/p_a	r_1/n_1	r/n	$E(N\|p_0)$	α	β	γ^L	r_i/i	$E(N\|p_0)$	α	β
0.10/0.30	1/12	5/35	19.84	0.0977	0.0986	−4.37	0/11 1/16 2/20 3/23 4/26	17.68	0.0990	0.0971
	1/16	4/25	20.37	0.0951	0.0970	−6.95	0/13 1/17 2/21 3/23 4/25	19.03	0.0936	0.0991
0.15/0.30	3/23	11/55	37.73	0.0998	0.0993	−0.53	0/11 1/17 2/22 3/27 4/31 5/35 6/39 7/43 8/47 9/51 10/55 11/58 12/62	32.41	0.0922	0.0997
	5/34	11/53	41.65	0.0867	0.0996	−7.53	0/18 1/23 2/28 3/31 4/35 5/38 6/41 7/44 8/46 9/49 10/51 11/53	38.54	0.0854	0.0999
0.20/0.35	5/27	16/63	43.61	0.0999	0.0981	−0.27	0/10 1/15 2/19 3/23 4/27 5/31 6/34 7/37 8/41 9/44 10/47 11/51 12/54 13/57 14/60 15/63 16/66 17/70	36.18	0.0994	0.0993
	6/33	15/58	45.49	0.0992	0.0997	−7.04	0/15 1/21 2/25 3/28 4/32 5/35 6/38 7/40 8/43 9/45 10/48 11/50 12/52 13/55 14/57 15/58	42.14	0.0985	0.0989
0.20/0.40	3/17	10/37	26.02	0.0948	0.0967	−0.80	0/10 1/12 2/15 3/19 4/22 5/25 6/28 7/30 8/33 9/36 10/38	21.75	0.0965	0.0996
	3/19	10/36	28.26	0.0861	0.0976	−5.88	0/11 1/15 2/18 3/21 4/24 5/26 6/29 7/31 8/33 9/35 10/36	24.84	0.0847	0.0995
0.30/0.45	9/30	29/82	51.38	0.0990	0.0995	−0.55	0/10 1/12 2/15 3/18 4/21 5/24 6/27 7/29 8/32 9/34 10/37 11/40 12/42 13/45 14/47 15/49 16/52 17/54 18/57 19/59 20/62 21/64 22/66 23/69 24/71 25/73 26/76 27/78 28/80	41.55	0.0980	0.0996
	16/50	25/69	56.01	0.0998	0.0984	−6.21	0/11 1/16 2/19 3/22 4/25 5/28 6/31 7/33 8/36 9/38 10/40 11/42 12/45 13/47 14/49 15/51 16/53 17/55 18/57 19/59 20/61 21/63 22/64 23/66 24/68 25/69	47.64	0.0986	0.0994
0.30/0.50	7/22	17/46	29.89	0.0974	0.0951	−0.10	1/10 2/12 3/15 4/17 5/20 6/22 7/25 8/27 9/29 10/31 11/34 12/36 13/38 14/40 15/42 16/45 17/47	24.24	0.0958	0.0980
	7/28	15/39	34.99	0.0943	0.0999	−4.30	0/10 1/11 2/14 3/17 4/19 5/21 6/23 7/26 8/28 9/30 10/31 11/33 12/35 13/37 14/38 15/40	25.75	0.0998	0.0998
0.40/0.60	7/18	22/46	30.22	0.0952	0.0996	−0.88	2/10 3/12 4/14 5/16 6/18 7/20 8/22 9/24 10/26 11/28 12/29 13/31 14/33 15/35 16/37 17/38 18/40 19/42 20/44 21/45 22/47	24.53	0.0972	0.0987
	11/28	20/41	33.84	0.0951	0.0991	−8.05	0/10 1/11 2/14 3/16 4/18 5/20 6/21 7/23 8/25 9/26 10/28 11/29 12/31 13/32 14/34 15/35 16/36 17/38 18/39 19/40 20/41	29.59	0.0945	0.1000
0.60/0.80	6/11	26/38	25.38	0.0970	0.0958	−0.22	4/10 5/11 6/12 7/13 8/15 9/16 10/18 11/19 12/20 13/22 14/23 15/24 16/25 17/27 18/28 19/29 20/31 21/32 22/33 23/35 24/36 25/37	19.58	0.0999	0.0988
	18/27	24/35	28.47	0.0965	0.0997	−1.89	4/10 5/11 6/12 7/14 8/15 9/17 10/18 11/19 12/20 13/22 14/23 15/24 16/26 17/27 18/28 19/29 20/31 21/32 22/33 23/34 24/35	19.95	0.0986	0.0992

TABLE 9 Additional comparisons with two-stage three-outcome designs (Sargent et al.) for Example 4

	Two-stage three-outcome design (Sargent et al.)							CF three-outcome design without early efficacy boundary							
	Optimal and Minimax							Optimal and Minimax							
p_0/p_a	r_1/n_1	$r, s/n$	$E(N\|p_0)$	α	β	η	π	γ^L	r_i/i	s_n/n	$E(N\|p_0)$	α	β	η	π
0.10/0.30	1/13	3,5/22	16.41	0.059	0.100	0.850	0.820	−1.61	0/10 1/14 2/19 3/22	5/22	15.36	0.059	0.099	0.822	0.851
	1/16	3,5/21	18.43	0.052	0.089	0.851	0.801	−9.49	0/13 1/16 2/19 3/21	5/21	17.24	0.052	0.093	0.800	0.854
0.15/0.30	2/17	8,11/46	30.93	0.067	0.099	0.802	0.820	−2.21	0/11 1/17 2/21 3/26 4/29 5/33 6/36 7/40 7/41	10/41	26.97	0.073	0.099	0.812	0.803
	2/22	7,9/37	31.93	0.092	0.100	0.823	0.821	−8.06	0/16 1/21 2/25 3/28 4/31 5/33 6/36 7/37	9/37	30.27	0.092	0.100	0.822	0.823
0.20/0.35	4/22	11,14/48	33.89	0.077	0.099	0.800	0.815	−0.16	0/10 1/14 2/18 3/22 4/26 5/29 6/32 7/36 8/39 9/42 10/46 11/49	14/49	30.17	0.088	0.099	0.837	0.803
	3/22	11,13/45	37.36	0.098	0.099	0.832	0.840	−2.92	0/11 1/16 2/19 3/23 4/26 5/30 6/32 7/35 8/38 9/41 10/43 10/44	13/44	30.20	0.083	0.100	0.809	0.802
0.20/0.40	2/13	7,9/29	20.97	0.100	0.093	0.817	0.856	−2.73	0/10 1/12 2/15 3/18 4/20 5/23 6/25 6/27	9/27	18.56	0.071	0.099	0.806	0.806
	2/16	6,8/24	21.19	0.089	0.099	0.813	0.807	−9.07	0/11 1/15 2/17 3/19 4/21 5/23 6/24	8/24	19.80	0.089	0.100	0.807	0.815
0.30/0.45	6/23	18,21/53	39.80	0.082	0.098	0.806	0.807	−1.37	0/10 1/12 2/15 3/18 4/20 5/23 6/26 7/28 8/31 9/33 10/36 11/38 12/41 13/43 14/45 15/47 16/50 17/52 18/54	21/54	34.49	0.096	0.099	0.833	0.806
	14/43	17,20/50	45.04	0.084	0.092	0.802	0.801	−6.04	0/11 1/15 2/18 3/21 4/23 5/26 6/28 7/30 8/33 9/35 10/37 11/39 12/41 13/43 14/44 15/46 16/48 17/50	20/50	37.46	0.084	0.095	0.800	0.805
0.30/0.50	3/12	12,15/35	23.67	0.067	0.100	0.812	0.812	−1.84	1/10 2/12 3/15 4/17 5/19 6/21 7/23 8/25 9/28 10/29 10/30	13/30	20.12	0.081	0.099	0.809	0.804
	6/20	11,14/32	24.70	0.067	0.084	0.803	0.803	−1.84	1/10 2/12 3/15 4/17 5/19 6/21 7/23 8/25 9/28 10/29 10/30	13/30	20.12	0.081	0.099	0.809	0.804
0.40/0.60	8/19	16,19/37	24.99	0.093	0.098	0.800	0.851	−0.13	2/10 3/11 4/14 5/15 6/17 7/19 8/21 9/23 10/25 11/27 12/28 13/30 14/32	17/32	20.04	0.087	0.100	0.818	0.800
	7/20	14,16/30	25.84	0.097	0.100	0.827	0.824	−7.85	1/10 2/13 3/15 4/16 5/18 6/20 7/21 8/23 9/24 10/25 11/27 12/28 13/29 14/30	16/30	23.54	0.097	0.100	0.824	0.827
0.60/0.80	7/13	16,18/24	19.32	0.095	0.097	0.814	0.809	−0.45	5/10 6/12 7/13 8/15 9/16 10/17 11/18 12/20 13/21 14/22 15/24 16/25 17/26 17/27	20/27	16.71	0.091	0.099	0.831	0.817
	7/13	16,18/24	19.32	0.095	0.097	0.814	0.809	−4.43	4/10 5/11 6/12 7/14 8/15 9/16 10/17 11/19 12/20 13/21 14/22 15/23 16/24	18/24	17.08	0.096	0.098	0.810	0.816

We obtain our optimal two-outcome CF procedure based on confidence function (8) and compare with the above designs. We set $\gamma^U = -\infty$ so that no early stopping for efficacy is possible and restrict $\gamma^L \in [-10, 0]$ when searching for a procedure that minimizes $E(N|\, p_0)$.

Our optimal CF design corresponds to $\gamma^L = -0.80$ with 38 as the maximum number of patients. The expected number of patients under H_0 is $E(N|\, p_0) = 21.75$, which is lower than that of the PP and the Simon designs. See Table 5 for additional information. Boundaries and probabilities of stopping for each analysis under H_0 and under H_a are shown in Table 6. We also search for the CF design that has the smallest maximum sample size n. This minimax CF design corresponds to $\gamma^L = -5.88$ with $n = 36$ and $E(N|\, p_0) = 24.84$. The boundary for rejecting H_a at the final analysis is $r_{36} = 10$.

Example 3. *A group sequential two-outcome procedure without early stopping for efficacy.*

Consider the same study setting as in Example 2. Suppose due to operational considerations, we prefer analyzing the data after every five additional patients following the first analysis at 10 patients. Of course, we always plan to analyze the study at the planned maximum number of patients even if it is not a multiple of 5. We derive such a study design based on confidence function (8).

For illustrative purposes, we search $\gamma^L \in [-10, 10]$ by step size of 0.01. Our CF design that minimizes $E(N|\, p_0)$ requires $n = 43$ patients unless early stopping occurs. The average number of patients under H_0 is 22.37. See Table 7.

Example 4. *Additional Examples.*

In Tables 8 and 9, we provide additional examples comparing our two- and three-outcome CF designs based on Equation (8) with traditional two-stage two- and three-outcome procedures (i.e., Simon and Sargent et al.). As these traditional designs do not have an early efficacy boundary, for comparison, we restrict our CF procedures to those without an early efficacy boundary. The specified α and β values are 0.1. For three-outcome procedures, $\eta \geq 0.8$ and $\pi \geq 0.8$ are targeted. For optimal and minimax CF designs, we searched all procedures with $\gamma^L \in [-10, 0]$ using step size of 0.01 allowing for continual monitoring after the first 10 patients. In general, our CF designs compare favorably to the traditional two- and three-outcome procedures.

5 | DISCUSSION

Tumor response has been frequently used in oncology phase II clinical trials. Because, without effective treatment, spontaneous and substantial tumor shrinkage that qualifies as a response according to RECIST[10] is very rare, a concurrent randomized control may not be necessary. In the single-arm study setting, two- and three-stage, two- and three-outcome designs[1-4] have been widely used. Within the constraint of the number of stages, these designs are globally optimal with respect to the given loss function as they are obtained by searching among all valid procedures. Even though no particular requirement is imposed on the interim and final boundaries (e.g., a high level of evidence for rejecting H_a after stage 1), these designs are usually reasonable. When the number of stages increases, especially when continual monitoring is desired, searching for an optimal design among all valid procedures becomes computationally prohibitive due to the large number of potential designs that one needs to consider.

In this paper, we propose a two- and three-outcome Bayesian posterior probability design based on confidence functions that allows for continual monitoring with or without an early efficacy boundary. Our CF design is based on the frequentist hypothesis testing framework meeting error rate and power specifications, but decision boundaries for futility and/or efficacy are formulated according to Bayesian posterior probability that H_a or H_0 is not true. Instead of searching globally, we search for an optimal design among procedures corresponding to a family of confidence functions. By specifying appropriate confidence functions to use, we can also require higher levels of evidence for rejecting H_0 or H_a early. The required level of evidence for futility and/or efficacy at each analysis has a clear and intuitive interpretation. Through several examples, we compared our design with commonly used two-stage procedures by Simon[1]

and Sargent et al.[2] as well as Bayesian predictive probability-based designs by Lee et al.[7] We found that our design typically has a lower average patient number, $E(N|\ p_0)$.

Despite typically lower average study sizes, analyzing accumulating data continually during the conduct of a clinical study could present substantial operational challenges. When sequential analyses are overly burdensome, group sequential designs as described in Section 2.6 or traditional two- or three-stage procedures such as Simon and Sargent et al. may be considered. This is especially true when enrolment pause needs to be implemented for each analysis in order to limit patient enrolment overrun.

Our design, given that data can be analyzed sequentially, has some similarities to the sequential probability ratio test (SPRT),[11] which was originally developed for use in quality control studies in manufacturing. With SPRT, new data are collected and analyzed continually, without a maximum limit on sample size, until pre-defined boundaries are crossed. For clinical trials, however, we typically need to cap the maximum study size so that a decision whether to reject the null or alternative hypothesis is made when the maximum study size is reached. Additionally, SPRT has constant boundaries while for clinical trials, it is often desirable to require higher levels of evidence to stop early for futility or efficacy. With our proposed design, this is achieved by specifying a desired confidence function. Furthermore, a family of confidence functions can be used to derive an "optimal" design that minimizes a specific loss function.

In oncology, there is typically a time gap of at least a few weeks between start of treatment and tumor response. A patient needs to be followed for a minimum amount of time before becoming evaluable for response. Even though $E(N|\ p_0)$ is commonly used as a loss function for designing clinical studies, others could be considered. In practice, with continuous patient enrolment, by the time the i^{th} patient is evaluable, more patients may be already enrolled into the trial. Therefore, futility at that time would mean a cost of i patients plus any enrolment overrun. In this situation, a loss function taking potential patient overrun into consideration may be helpful. Additionally, for settings where an informative prior for p, say $g(p)$, is available at study planning, one may choose to minimize expected value of $E(N|\ p)$ under the prior distribution of p, $\int E(N|\ p)g(p)dp$. In the absence of any reliable prior information on p, it is advisable to obtain both the minimax and optimal designs and compare them before deciding on the design for a given study. If, for example, $E(N|\ p_0)$ for the minimax design is only fractionally higher than that of the optimal design and the optimal design requires many more patients in maximum, it may make good sense to choose the minimax design.

For any clinical trial with early stopping boundaries, if there is enrolment overrun at the time of early stopping, the final result including all enrolled patients might not always be consistent with the result at the time of early stopping, even though such an inconsistent outcome should be rather rare. However, decisions based on the pre-defined early stopping boundaries ensure that false positive and negative rates are controlled regardless of the "final" result. In the current setting of signal-generating phase II studies, any decision on future development should be made based on the totality of data available at the time of the decision.

Our CF design allows an optional early efficacy boundary in addition to a futility boundary. Even though there is no compelling ethical reason to stop a single-arm study early due to early signals of efficacy when every patient is on the promising therapy, an early boundary for efficacy can still be useful in certain situations. For example, one may wish to initiate a confirmatory trial as soon as early efficacy boundary is crossed in order to reduce the overall development time. We use the term early stopping for efficacy in this article loosely to mean early rejection of the null hypothesis and do not necessarily mean enrolment will be stopped. However, even though the type I and II error rates are strictly controlled for making a drug development go/no-go decision, the resulting point estimate of ORR may be biased after early stopping.[12] Therefore, even if the efficacy boundary is crossed, if one needs a more precise and unbiased estimate of ORR, the study should be allowed to continue until the planned maximum number of patients because there is no ethical concern to treat additional patients when the experimental therapy is promising.

Finally, when monitoring a single-arm phase II study continually, it is important to follow the natural order of patient enrolment. For the analysis at the i^{th} patient, the first, and only the first, i patients enrolled consecutively into the trial should be included. Without following such a predefined order, type I and II error rates may become ambiguous and not controlled. For example, if we include the first i patients who complete study treatment, the analyzed group may be enriched with patients who progress and discontinue treatment earlier, therefore biasing the data and results.

An executable Java application with a graphical user interface has been developed for obtaining the proposed study designs and it is available from the author upon request.

CONFLICT OF INTEREST

The author declares that there is no conflict of interest that could be perceived as prejudicing the impartiality of the research reported.

DATA AVAILABILITY STATEMENT

Data sharing not applicable to this article as no datasets were generated or analysed during the current study.

ORCID

Minghua Shan https://orcid.org/0000-0002-6858-9769

REFERENCES

1. Simon R. Optimal two-stage designs for phase II clinical trials. *Control Clin Trials*. 1989;10:1-10.
2. Sargent DJ, Chan V, Goldberg RM. A three-outcome design for phase II clinical trials. *Control Clin Trials*. 2001;22:117-125.
3. Chen K, Shan M. Optimal and minimax three-stage designs for phase II oncology clinical trials. *Contemp Clin Trials*. 2008;29:32-41.
4. Ensign LG, Gehan EA, Kamen DS, Thall PF. An optimal three-stage design for phase II clinical trials. *Stat Med*. 1994;13:1727-1736.
5. Green SJ, Dahlberg S. Planned versus attained design in phase II clinical trials. *Stat Med*. 1992;11:853-862.
6. Koyama T, Chen H. Proper inference from Simons two–stage designs. *Stat Med*. 2008;27(16):31453154.
7. Lee JJ, Liu DD. A predictive probability design for phase II cancer clinical trials. *Clin Trials*. 2008;5(2):93-106.
8. Hwang IK, Shih WJ, De Cani JS. Group sequential designs using a family of type I error probability spending functions. *Stat Med*. 1990;9 (12):1439-1445.
9. Jung S, Lee T, Kim K, George SL. Admissible two-stage designs for phase II cancer clinical trials. *Stat Med*. 2004;23:561-569.
10. Eisenhauer EA, Therasse P, Bogaerts J, et al. New response evaluation criteria in solid tumours: revised RECIST guideline (version 1.1). *Eur J Cancer*. 2009;45(2):228-247.
11. Wald A. Sequential tests of statistical hypotheses. *Ann. Math. Stat.* 1945;16(2):117-186.
12. Whitehead J. On the bias of maximum-likelihood-estimation following a sequential test. *Biometrika*. 1986;73(3):573-581.

How to cite this article: Shan M. A confidence function-based posterior probability design for phase II cancer trials. *Pharmaceutical Statistics*. 2021;20:485–498. https://doi.org/10.1002/pst.2089