Published in final edited form as:

Pharm Stat. 2019 November ; 18(6): 700–713. doi:10.1002/pst.1965.

A two-stage phase II clinical trial design with nested criteria for early stopping and efficacy

Michelle DeVeaux, Michael Kane, Wei Wei, Daniel Zelterman

Department of Biostatistics School of Epidemiology and Public Health, Yale University, New Haven, Connecticut

Abstract

We propose a two-stage design for a single arm clinical trial with an early stopping rule for futility. This design employs different endpoints to assess early stopping and efficacy. The early stopping rule is based on a criteria determined more quickly than that for efficacy. These separate criteria are also nested in the sense that efficacy is a special case of, but usually not identical to, the early stopping endpoint. The design readily allows for planning in terms of statistical significance, power, expected sample size, and expected duration. This method is illustrated with a phase II design comparing rates of disease progression in elderly patients treated for lung cancer to rates found using a historical control. In this example, the early stopping rule is based on the number of patients who exhibit progression-free survival (PFS) at 2 months post treatment follow-up. Efficacy is judged by the number of patients who have PFS at 6 months. We demonstrate our design has expected sample size and power comparable with the Simon two-stage design but exhibits shorter expected duration under a range of useful parameter values.

Keywords

curtailed sampling; expected sample size; expected trial duration; minimax design

1 | INTRODUCTION

A fundamental design feature of the two-stage phase II clinical trial is the convenient early stopping rule preventing us from exposing excessive numbers of patients to a possibly inferior treatment. In the traditional Simon two-stage designs, we use the same criteria for the early stopping endpoint as for the efficacy endpoint.¹ The outcomes of patients enrolled in stage 1 are observed before enrolling additional subjects in stage 2. The criteria for advancing from stage 1 to stage 2 are based on a (statistically) large number of patients achieving the same favorable endpoint in stage 1 as will be later examined for efficacy of treatment at the conclusion of stage 2. An ideal endpoint for such a design is one that can be assessed quickly. A good example of endpoint in cancer treatment might include tolerability or response to therapy, as determined a few weeks after treatment, verified by a CT scan. It is often not possible to choose a quickly observable endpoint. The motivation for our present

work is to improve a trial design in which the efficacy determination can only be made after a lengthy follow-up time on each patient.

Our motivating problem came about in a study of a novel drug combination treatment of lung cancer in geriatric patients for which the efficacy was chosen to be the rate of 6-month progression-free survival (PFS). This criteria would create a long delay between stage 1 and stage 2 as we wait to observe the outcomes of the stage 1 patients before enrolling additional patients into stage 2. Using traditional designs, there could be up to a 6-month delay in accrual until the outcomes of all stage 1 patients become available. One solution to this delay would be to enroll all patients without an early stopping rule. Such a design eliminates any safeguards of exposing excessive numbers of patients to an inferior or harmful treatment. Another solution would be to continue accruing patients into stage 2 while waiting for those in Stage 1 to achieve their endpoints. This approach also has the potential of exposing too many patients to a possibly inferior treatment.

Our proposed solution is to use an interim endpoint for early stopping different from the efficacy endpoint. We want these separate endpoints to be nested in the sense the efficacy endpoint is a special case of that of futility. That is, the efficacy endpoint cannot be achieved if the patient fails the futility endpoint. In our specific application in section 5, the efficacy measure is the rate of 6-month PFS, and 2-month PFS rate is the criteria for early stopping. The estimates of 2- and 6-month historical PFS rates are based on a published Kaplan-Meier survival curve $\hat{S}(t)$ estimating PFS times. The published survival curve $\hat{S}$ provides a standard-of-care comparison and provides us with the flexibility to choose any pair of PFS times and the corresponding survival rates.

The large volume of literature written on two-stage designs attests to their popularity in clinical trial practice. We cite only some of the literature here. The original idea of the two-stage design for phase II trials is attributed to R. Simon,¹ with other developments building on the original idea.^{2–8} There are review articles covering phase II designs and, more generally, issues in clinical trials.^{9–13} The present work is concerned with separate endpoints for futility and efficacy. Designs with separate endpoints for safety and efficacy have been introduced,^{14–18} but those criteria are not nested as is proposed here. We will also consider curtailed sampling schemes^{19–24} allowing for early decisions to be made before all of the data have been observed.

In section 2, we provide the necessary notation and derive some of the basic mathematical formulæ. In section 3, we consider curtailed sampling allowing early decisions to be made before the data from the entire sample are available to us. In section 4, we derive expressions to compute the expected trial duration. In section 5, we provide practical guidance using the parameters of the proposed lung cancer trial as an example. In section 6, we compare designs for a range of different parameter values. The details of the computations for expected sample size, minimax sampling, and expected trial duration are given in the appendices.

2 | NOTATION AND BASIC RESULTS

We develop the proposed design by considering rates of overall survival, rates of PFS, or other outcomes measured at two follow-up times t_1 – t_2 . Overall survival or PFS at two different times is the most obvious example of a pair of nested criteria fitting the framework of this design.

Typical planning for the phase II studies proposed here begins with a published Kaplan-Meier survival (or PFS) estimate $\hat{S}(t)$, which we use as a benchmark for the established standard of care. The new treatment under consideration will be compared with this estimate of a historical control. The estimated rates of survival, $\hat{p}_i = \hat{S}(t_i)$, ($i = 1, 2$) obtained from this survival curve provide us with a pair of parameters needed to design our study. We will omit the carets on $\hat{p}_1$ and $\hat{p}_2$, but the reader should keep in mind these historical control rates are usually estimates, as is typically the case with many phase II studies.

The null hypothesis we wish to test is H_0 : PFS at time t_2 is equal to p_2 . The alternative hypothesis is this rate is greater than p_2 . The survival rate p_1 at time t_1 is a futility parameter, and we will terminate the trial early if the observed PFS at time t_1 is statistically lower than p_1 . These rates satisfy $0 \leq p_2 \leq p_1 \leq 1$. In the traditional Simon two-stage design,¹ we have $t_1=t_2$ and $p_1=p_2$ because the futility and efficacy criteria are identical. We will discuss these parameters again and provide a specific example in section 5.

We will need to identify study parameters for a design, terminating the clinical trial early if we have statistical evidence the PFS fraction at time t_1 is less than p_1 . Let n_1 denote the sample size of patients enrolled in stage 1 of the study. Let the random variable X_1 denotes the number of these patients who are still progression free after being followed for a duration of time t_1 .

If patients have independent and homogeneous response rates, then X_1 behaves as a binomial random variable with parameters n_1 and p_1 . For an a priori fixed critical value r_1 , if $X_1 \geq r_1$, then the trial will continue from stage 1 into stage 2. The design parameter r_1 will need to be determined as part of the clinical trial planning process. The X_1 successful patients in stage 1 will continue to be followed at least as far as time t_2 if we decide to proceed into stage 2. If $X_1 < r_1$ then the trial will terminate early without continuing into stage 2.

In stage 2, we enroll additional patients, follow each of them for time t_2 , and use their responses to test the null hypothesis for efficacy. Specifically, in stage 2, we will enroll an additional n_2 patients and follow each for a minimum period of time t_2 where $t_2 \geq t_1$. Let the random variable X_2 denote the number of successes among these stage 2 patients. The distribution of X_2 is binomial with index parameter n_2 and probability of success p_2 , again assuming the historical rates under the null hypothesis.

In addition to the X_2 successes in stage 2, there will also be a random variable, denoted by X_{12} , of successes at time t_2 who were initially enrolled in stage 1. That is, X_{12} is the conditional number of stage 1 patients who are progression free at time t_2 given they were

progression free at time t_1 . The conditional distribution of X_{12} given X_1 is binomial with index X_1 and probability parameter p_2/p_1 . The distribution of X_2 is independent of both X_1 and X_{12} .

At the conclusion of stage 2, we will reject the null hypothesis and declare the new treatment successful if the total number of successes ($X_{12}+X_2$) at time t_2 exceeds a critical value r_2 determined a priori. Figure 1 is a useful representation of the timeline for this design. Given these notations, we can proceed with some basic derivations for evaluating the operating characteristics of the proposed clinical trial design of this form.

The probability of early termination is

$$\Pr[\text{Early termination}] = \Pr[X_1 < r_1], \quad (1)$$

where X_1 has a binomial distribution with index n_1 and probability parameter p_1 .

From this expression, the expected sample size is

$$\text{Expected sample size} = n_1 + n_2(1 - \Pr[\text{Early termination}]). \quad (2)$$

In the following section, we consider curtailed sampling in which the clinical trial can end as soon as any statistical endpoint has been reached. Under curtailed sampling, we obtain a lower expected sample size than given in (2).

Let us next derive an expression for the probability we reject the null hypothesis H_0

$$\Pr[\text{reject } H_0] = \Pr[X_1 \geq r_1 \text{ and } X_{12} + X_2 \geq r_2].$$

That is, we reject the null hypothesis if the trial does not terminate early in stage 1, and there are a sufficient number of successes in stage 2.

We first condition on the number of successes X_1 in stage 1

$$\Pr[\text{reject } H_0] = \sum_{x_1} \Pr[X_1 = x_1] \Pr[X_{12} + X_2 \geq r_2 | X_1 = x_1].$$

Since X_{12} and X_2 are independent, we can write the distribution of their sum as a convolution

$$\Pr[\text{reject } H_0] = \sum_{x_1} \Pr[X_1 = x_1] \sum_{x_{12}} \Pr[X_{12} = x_{12} | x_1] \Pr[X_2 \geq r_2 - x_{12}].$$

This gives us

$$\Pr[\text{reject } H_0] = \sum_{x_1=0}^{n_1} \Pr[X_1 = x_1] \sum_{x_{12}=0}^{x_1} \Pr[X_{12} = x_{12} | x_1] \sum_{x_2=0}^{n_2} \Pr[X_2 = x_2] \quad (3)$$

$$\max(r_1, r_2 - n_2) \quad \max(0, r_2 - n_2) \quad \max(0, r_2 - x_{12})$$

We also have

$$\text{Power} \leq 1 - \Pr[\text{Early termination | alternative hypothesis}] \quad (4)$$

because we cannot reject the null hypothesis if we decide to terminate the clinical trial early. This bound proves to be critical in choosing a suitable design, as we show in section 5. The probability of early stopping depends on p_1 , which may or may not be different under the alternative hypothesis.

For specified values of (p_1, p_2) , we find values of the design parameters (n_1, n_2, r_1, r_2) , so the probability in (3) does not exceed a specified significance level $\alpha > 0$. These parameters allow us to calculate the probability of early termination and the bound on the expected sample size given at (2). Table 1 summarizes a variety of designs for the lung cancer example.

To control for the significance level α , we first note (3) is an increasing function of p_1 . That is, the lower the probability of early stopping, the greater the significance level. The significance level in (3) is then maximized for $p_1=1$ corresponding to a binomial design with with index n_1+n_2 , probability of success p_2 , and no early stopping for the trial. Similarly, the choice of r_2 is the same as the value used in a binomial design with no early stopping rule.

The power of the design is the probability of rejecting the null hypothesis. The probability in (3) is a general statement, valid for any values of (p_1, p_2) satisfying $0 \leq p_2 \leq p_1 \leq 1$. In addition to describing the statistical significance at the null hypothesis, we can consider other parameter values as well and use this same expression to determine the power at other choices of (p_1, p_2) . The designs proposed in section 5 aim to test an increase in the value of the efficacy parameter p_2 while holding the futility parameter p_1 fixed under both hypotheses. We can also use (3) to describe power against other alternative hypotheses where p_1 as well as p_2 are allowed to vary between hypotheses. Figure 2 illustrates power for one design in the lung cancer example for all choices of (p_1, p_2) .

3 | CURTAILED SAMPLING

We introduce curtailed sampling into this design when we allow for early decisions to be made once sufficient observations justify any of the three statistical endpoints (stop early, accept, or reject the null hypothesis) to be decided. Let Y_1 and Y_2 denote the random

variables representing the sample sizes needed to achieve critical statistical endpoints in stage 1 and stage 2, respectively. The expectation of $Y_1 + Y_2$ is derived in Appendix A.

Let us describe some properties of this sampling scheme. We decide to continue to stage 2 as soon as r_1 successes are observed in stage 1. Similarly, we can also terminate early as soon as $n_1 - r_1 + 1$ failures are observed in stage 1. That is, stage 1 ends as soon as we observe either r_1 successes or else $n_1 - r_1 + 1$ failures. The range of Y_1 is then

$$\min(r_1, n_1 - r_1 + 1) \leq Y_1 \leq n_1.$$

The marginal distribution of Y_1 is a stopped negative binomial (SNB) and derived by us in DeVeaux et al.²⁵ The SNB(p, s, t) is the distribution of the number of independent Bernoulli(p) random variables necessary in order to observe either s successes or else t failures. At the point, we decide to continue from stage 1 to stage 2, there will be r_1 successes and $Y_1 - r_1$ failures in stage 1. Details about the distribution of Y_1 appear in Appendix A.

In studies not terminating early under curtailed sampling, we have $X_1 = r_1$, and the distribution of X_{12} is binomial with index r_1 and probability parameter p_2/p_1 . Under curtailed sampling, we will assume we observe all the stage 2 outcomes of the X_{12} patients before we observe the outcome of the first patient enrolled in stage 2.

The stage 2 curtailed sample size Y_2 may take advantage of the early decision in stage 1 to continue into stage 2. We will make use of the $n_1 - Y_1$ extra patients available to us in stage 2, without exceeding the maximum planned sample size of $n_1 + n_2$. The range of Y_2 is then

$$\max(0, r_2 - r_1) \leq Y_2 \leq n_2 + n_1 - Y_1.$$

The distribution of X_2 given Y_1 is binomial with index $n_2 + n_1 - Y_1$ and probability parameter p_2 .

The probability of rejecting the null hypothesis under curtailed sampling in the present design will be the same as the probability under conventional (noncurtailed) sampling (3). The critical values r_1 and r_2 will be identical under both sampling procedures. The expectation of $Y_1 + Y_2$ is detailed in Appendix A. Appendix B derives the probability

$$\Pr[Y_1 + Y_2 = n_1 + n_2] \tag{5}$$

of the maximum curtailed sample size required to complete the trial. This probability is needed to plan a minimax design.

4 | TRIAL DURATION

One motivation for the present work is to improve trial designs where the efficacy endpoint involves lengthy follow-up time. Our goal is to reduce the suspension in patient accrual occurring at the end of stage 1 in Simon's two-stage design and shorten the overall length of

this type of trial. Trial duration includes the time for patient accrual and their follow-up. In Appendix C, we derive the expected duration of the nested design under curtailed sampling. The expected duration of Simon's design under both traditional binomial and curtailed sampling is given in Appendix D. Case and Morgan²⁶ examine the duration of noncurtailed two-stage designs.

Let λ denote the average rate of patient enrollment per unit of t_1 and t_2 time, usually taken as per month. The expected duration of the nested design, without curtailed sampling, is

$$n_1/\lambda + t_1 + (1 - \text{Pr}[\text{early termination}])(n_2/\lambda + t_2),$$

where $\text{Pr}[\text{early termination}]$ is given at (1).

The expected duration of Simon's two-stage design with n_1 stage 1 observations and corresponding critical value r_1 is

$$n_1/\lambda + t_2 + (1 - \text{Pr}[\text{early termination}])(n_2/\lambda + t_2) \quad (6)$$

again without curtailed sampling, where $\text{Pr}[\text{early termination}] = \text{Pr}[B < r_1]$ and B is distributed as $\text{Binomial}(n_1, p_2)$.

In Appendix C, we use the distribution of the sample size under curtailed sampling to derive an expression for the expected duration of the nested design and in Appendix D for Simon's design.

5 | DESIGN GUIDANCE

In this section, we provide guidance on how to select a useful and practical design for the lung cancer example in the Introduction and compare it with the Simon two-stage design. We presume the most important decision in planning this type of clinical trial is the choice of sample size, usually determined by the costs. We will not address this issue. In other words, we assume the overall sample size, $n_1 + n_2$, is fixed and determined in another manner. Statistical significance is usually taken as $\alpha = .1$ for the smaller studies we have in mind. The choice of (p_1, p_2) or equivalently, (t_1, t_2) , could be decided on the expected duration of the trial, as discussed in section 4. There may be practical considerations for the choices of t_1 and t_2 , such as timing of periodic CT scans for evaluation of disease progression.

In light of (4), we propose limiting the probability of early termination in the nested design to values in the interval (.05, .2) and choose the largest value of r_1 such that the probability of early stopping does not exceed this limit. Our principal concern in stage 1 is testing whether the futility endpoint rate p_1 is substantially lower than the historical rate in which case the probability of stopping the trial early will be higher than our set limit.

There is wide latitude in allocating sample sizes (n_1, n_2) to the two stages. Following Simon,¹ the optimal design choices of (n_1, n_2) are those minimizing the expected sample size subject to fixed values of $n_1 + n_2$. Similarly, we can fix the overall sample size $n_1 + n_2$ and choose (n_1, n_2) to minimize the probability (5) requiring the maximum sample size. Such minimax

nested trials can be designed using the expression for this probability derived in Appendix B. Finally, there are balanced designs⁸ with $n_1=n_2$.

To demonstrate the use of the nested design, let us develop the motivating example from the Introduction. In this illustration, we restrict designs to those with a maximum of $n_1+n_2=36$ patients to be enrolled. The figure of 36 is in line with the size of designs for phase II studies we regularly conduct at the Yale Cancer Center. Typical of such phase II studies, we also use a statistical significance of $\alpha=.1$. We consider the futility parameter $p_1=.8$ to model the PFS rate at follow-up time $t_1=2$ months in stage 1. We also assume $p_2=.2$ to describe the longer-term PFS rate at $t_2=6$ months in stage 2. These times and rates are suggested by a published survival curve we use as the basis for our historical control values.²⁷ For these parameter values, the optimal and minimax nested designs both have $n_1=6$ and $n_2=30$. We refer to the minimax and optimal design as “A” in Tables 1 and 2. Design “E” is the balanced design⁸ where $n_1=n_2=18$.

Table 1 illustrates the power for an alternative hypothesis of $p_1=.8$ and $p_2=.4$ for each of the designs listed. Specifically, the futility rate p_1 is the same under both the null and alternative hypotheses but only the efficacy rate p_2 changes. Designs A to E have negligible differences in power, so the choice of which to use should incorporate our relative concerns for futility and the probability of inadvertently terminating the trial early for a useful therapy.

Our planning may also include varying both p_1 and p_2 between hypotheses. In this case, Figure 2 provides a comprehensive display of power for the minimax/optimal nested design A as an example. The null hypothesis in this case ($p_1=.8$ and $p_2=.2$) is indicated. In this figure, we can consider the power at any other pair of parameter values satisfying $0 \leq p_2 \leq p_1 \leq 1$. We see greater changes in power can be achieved by varying p_2 than for p_1 .

To complete a discussion of the designs for clinical trials, let us add three designs not of the nested form described in section 2. These three designs (binomial, Simon optimal, and Simon minimax) do not have separate endpoints for stopping and efficacy. The binomial design in Tables 1 and 2 has no early stopping and a fixed sample size of 36 patients. If 11 or more patients exhibit PFS after time t_2 (which occurs with probability .2 under the null hypothesis), then we reject this hypothesis at significance level .0889. The Simon optimal design has a maximum of 37 patients, and the Simon minimax design has 36 patients.

In terms of the notation in section 2, Simon’s two-stage design¹ is a special case of the nested designs with $t_1=t_2$ and $p_1=p_2$. Simon’s design uses the same criteria for early stopping and efficacy. Specifically, the traditional Simon design counts all the stage 1 successes towards rejecting the null hypothesis and consequently, should have greater power than the nested designs proposed. But as we see in Table 1, this improvement is modest. The binomial design also has the highest power of all designs considered but lacks any early stopping rule. Under the null hypothesis, Simon’s designs with comparable sample sizes will have greater probability of early termination than nested designs, again, because early stopping and efficacy decisions are based on the same outcomes.

Table 2 displays the expected sample sizes under curtailed sampling. We see similar expected sample sizes in designs A to E as compared with Simon’s designs with curtailed

sampling under both the null and alternative hypotheses. Introducing curtailed sampling rules reduces the trial's expected sample size, as shown by values in Tables 1 and 2.

Figure 3 demonstrates a large difference in expected duration of the trials. Near the null hypothesis, the Simon optimal design benefits from a large early stopping rule, but under the alternative, the nested design shows a large reduction in expected duration. This difference increases with the accrual rate λ . Slow accrual means follow-up times is less important in determining the expected trial duration. Nested designs B to E are inferior to the optimal/minimax design A under almost every criteria illustrating the values of (n_1, n_2) are important in choosing a good design.

6 | DISCUSSION

We constructed Table 3 in order to generalize from the parameter values in the lung cancer example of the Introduction to other hypotheses. This table uses most of the hypothesis tests on values of the p parameters considered in Chi and Chen.²¹ For each pair of null and alternative hypotheses, we began by identifying the corresponding Simon optimal design with significance level .1 and power .9. We next found the Binomial and nested optimal designs with the same maximum sample sizes. Finally, we calculated the expected duration for follow up times 2 and 6 months in the nested design and 6 months for the Binomial and Simon optimal designs. The expected durations were found assuming accrual rates of $\lambda=2$ and $\lambda=4$ patients per month.

Over all designs and hypotheses, the Simon optimal always comes closest to the nominal significance level of .1. For null hypotheses with small values of the p -parameters, both the binomial and nested optimal designs can be much smaller than this value, in part, because the sample sizes were optimized for the Simon design.

The power for the Simon optimal designs was specified to be at least .9. The binomial designs all have power comparable with this value. The nested designs always have lower power at least .8, except in one case. Then again, the sample sizes were optimized for the Simon designs.

The expected curtailed sample sizes for the Simon optimal designs are generally the smallest of the three designs in Table 3. The values given here all agree with those appearing in Table 2 of Chi and Chen.²¹ Under the null hypotheses, the Simon optimal design has smaller expected sample size than that of the nested optimal design because of the greater probability of early termination. Under the alternative hypothesis, however, the expected sample sizes of these two designs are more comparable.

The expected duration for the nested optimal is better than that of the Simon optimal for the larger accrual rate $\lambda=4$ we considered. Some of these differences can be quite large as we also see in Figure 3. When accrual rates are low then the Simon optimal design has shorter expected duration near the null hypothesis because of a high probability of terminating early. Further, with low accrual rates, the time for accrual will dominate the follow-up times. Noncurtailed sampling and Simon's design generally require longer trial durations but also contribute additional safety data.

For readers wishing to examine other conditions, our R code for all computations is available on the journal webpage for this article.

ACKNOWLEDGEMENT

This research was supported by grants P30 CA016359; R01 ES005775; P50CA196530; R01ES019179; R35CA197574 awarded by the National Institutes of Health; and additional support from the Yale Comprehensive Cancer Center; and DOD-US Army Medical Research and Material Command. M. DeVeaux is currently employed by Regeneron Pharmaceuticals, Tarrytown, NY. Our R code is available from the corresponding author. The authors thank Beth Nichols for her editorial assistance.

Funding information

National Institutes of Health, Grant/Award Numbers: P30 CA016359, R01 ES005775, P50CA196530, R01ES019179 and R35CA197574; Yale Comprehensive Cancer Center; DOD-US Army Medical Research and Material Command

APPENDIX A

EXPECTED CURTAILED SAMPLE SIZE

In this Appendix, we derive the distribution of the smallest sample size needed to reach any of the statistical endpoints of the nested clinical trial design under curtailed sampling. Similar expressions are obtained for curtailed sampling in the Simon two-stage design by Chi and Chen²¹ and in Appendix D.

If the trial terminates early, then $Y_2=0$. We then condition on early termination

$$\begin{aligned} E(Y_1 + Y_2) &= E(Y_1 | \text{early termination}) \Pr[\text{early termination}] \\ &+ E(Y_1 | \text{continue to stage 2}) (1 - \Pr[\text{early termination}]) \\ &+ E(Y_2 | \text{continue to stage 2}) (1 - \Pr[\text{early termination}]). \end{aligned} \quad (\text{A1})$$

Begin with the expectation of Y_1 . In trials terminating early, we stop sampling at the point we have observed $n_1 - r_1 + 1$ stage 1 failures and fewer than r_1 successes. Then Y_1 behaves as a truncated negative binomial random variable with

$$\Pr[Y_1 = y_1 | \text{early termination}] \propto \binom{y_1 - 1}{n_1 - r_1} p_1^{y_1}$$

defined for $y_1 = n_1 - r_1 + 1, \dots, n_1$.

In those trials continuing to stage 2, Y_1 behaves as a truncated negative binomial random variable counting the number of stage 1 samples until we observe $X_1 = r_1$ successes and at most $n_1 - r_1$ failures. The distribution of Y_1 is then

$$\Pr[Y_1 = y_1 | \text{continue to Stage 2}] \propto \binom{y_1 - 1}{r_1 - 1} (1 - p_1)^{y_1}, \quad (\text{A2})$$

defined for $y_1 = r_1, \dots, n_1$.

In studies not terminating early, the range of Y_2 is

$$\max(0, r_2 - r_1) \leq Y_2 \leq n_1 + n_2 - Y_1$$

because we can make use of the stage 1 sample size savings in stage 2.

For patients enrolled in stage 2, we stop sampling as soon as we observe either

$$s = \max(0, r_2 - x_{12})$$

successes or else

$$t = \max(0, n_2 + (n_1 - y_1) - (r_2 - x_{12}) + 1)$$

failures. Then Y_2 follows the SNB(p_2, s, t) distribution.²⁵ The expected value of Y_2 can be calculated numerically.

The distribution of Y_2 depends on the random variables Y_1 and X_{12} . Given the trial continues to stage 2, the expectation of Y_2 can be written as

$$\begin{aligned} E(Y_2 | \text{continue to stage 2}) = \\ \sum_{y_1=r_1}^{n_1} \Pr[Y_1 = y_1 | \text{continue}] \sum_{x_{12}=0}^{r_1} \Pr[X_{12} = x_{12}] E(Y_2 | y_1, x_{12}, \text{continue}), \end{aligned}$$

where $X_{12} \sim \text{Binomial}(r_1, p_2/p_1)$ and the distribution of Y_1 is given at (A2).

This concludes the necessary expressions to evaluate the expectation (A1). This expression has been confirmed by simulations, not reported here.

APPENDIX B

MINIMAX CURTAILED SAMPLING

We find the probability the full sample of size $n_1 + n_2$ is necessary when employing curtailed sampling in a nested design. Using the notation of Appendix A,

$$\Pr[Y_1 + Y_2 = n_1 + n_2] = (1 - \Pr[\text{early termination}]) \Pr[Y_1 + Y_2 = n_1 + n_2 | \text{continue}].$$

The probability of terminating early is given at (1) Next,

$$\Pr[Y_1 + Y_2 = n_1 + n_2 | \text{continue to Stage 2}] = \sum_{j=r_1} \Pr[Y_1 = j | \text{continue}] \Pr[Y_2 = n_1 + n_2 - j | Y_1 = j, \text{continue}]. \quad (\text{B1})$$

for summation in j over

$$r_1 \leq j \leq \min(n_1, n_1 + n_2 - 1).$$

The negative binomial distribution of the Y_1 trials needed to observe $r_1 > 0$ successes obeys

$$\Pr[Y_1 = j | \text{continue to Stage 2}] \propto \binom{j-1}{r_1-1} (1-p_1)^j$$

for $r_1 \leq j \leq n_1$.

The maximum sample size is achieved when the failure or success of the last patient enrolled determines whether we accept or reject the null hypothesis, respectively. Let X_2' be a random variable distributed as binomial with index $n_1 + n_2 - y_1 - 1$ and probability parameter p_2 . Under curtailed sampling, $X_{12} \sim \text{Binomial}(r_1, p_2/p_1)$. Then $X_{12} + X_2'$ describes the number of successes observed after the penultimate patient and

$$\begin{aligned} \Pr[Y_2 = n_1 + n_2 - j | Y_1 = j, \text{continue}] &= \Pr[X_{12} + X_2' = r_2 - 1] \\ &= \sum_i \Pr[X_{12} = i] \Pr[X_2' = r_2 - i - 1] \end{aligned}$$

which can be found numerically.

The range on i in this last summation is over the integer values in

$$\max\{0, r_2 + j - n_1 - n_2\} \leq i \leq \min\{r_1, r_2 - 1\},$$

and j is defined following (B1).

In the special case of $r_1 = 0$, there is zero probability of stopping early, and the probability of needing the maximum sample size is

$$\binom{n_1 + n_2 - 1}{r_2 - 1} p_2^{r_2 - 1} (1 - p_2)^{n_1 + n_2 - r_2}.$$

APPENDIX C

EXPECTED NESTED TRIAL DURATION

Let $\lambda > 0$ denote the enrollment rate of patients per unit of t_i times. Let Y_1 and Y_2 denote the curtailed sample sizes needed to achieve critical endpoints in stage 1 and stage 2, respectively. If the trial stops early, then the nested trial duration will be the enrollment time for the last patient plus follow-up time t_1 . Alternatively, if we observe r_1 successes at time t_1 , then the trial will continue to stage 2, and we will continue to follow these r_1 patients to time t_2 .

In stage 2, we stop sampling as soon as we observe either

$$s = \max(0, r_2 - x_{12}) \quad (C1)$$

successes or

$$t = \max(0, n_2 + (n_1 - y_1) - (r_2 - x_{12}) + 1) \quad (C2)$$

failures.

If we continue to stage 2 and either s or t take the value 0, then no additional observations are required to reach a statistical endpoint and Y_2 equals 0. If $X_{12} = r_2$ then $s=0$, and we can reject the null hypothesis because we have already observed a sufficient number of successes based on the stage 1 patients alone.

Similarly, if

$$r_2 - X_{12} \geq n_2 + n_1 - y_1 + 1,$$

then $t=0$, and we have already observed too many failures to be able to reach r_2 successes given the number of patients left to be enrolled in the trial. In this situation, we can accept the null hypothesis without enrolling any additional patients into stage 2. If either $s=0$ or $t=0$, then the duration of stage 2 is equal to $t_2 - t_1$, the additional follow-up time on the last patient enrolled in stage 1 needed to evaluate their efficacy criteria.

If the trial continues to the second stage, and s and t are both greater than 0, then additional patients need to be enrolled into the trial to reach a statistical endpoint. In this case, $Y_1 > 0$, and we assume the patient follow-up time will be t_2 . The expected duration of the nested trial design under curtailed sampling is

$$E(Y_1 + Y_2)/\lambda + t_1 + (1 - \Pr[\text{early termination}]) \times [(t_2 - t_1)\Pr[Y_2 = 0 | \text{continue to Stage 2}] + t_2\Pr[Y_2 > 0 | \text{continue to Stage 2}]],$$

where

$$\Pr[Y_2 = 0 | \text{continue to Stage 2}] = \sum_{y_1=r_1}^{n_1} \Pr[Y_1 = y_1 | \text{continue to Stage 2}] \times \sum_{x_{12}=0}^{r_1} \Pr[X_{12} = x_{12}] I\left\{s = 0 \cup t = 0\right\},$$

and s and t are given at (C1) and (C2), respectively.

An expression for $E(Y_1 + Y_2)$ is given in Appendix A. Conditional on values of y_1 and x_{12} , the probability $Y_2=0$ will be 1 if either s in (C1) or t in (C2) are 0. The probability $Y_2=0$ is 0 if s and t are both nonzero. Similarly, $\Pr[Y_2=0] = 0$ if $r_1 < r_2$.

APPENDIX D

SIMON'S DESIGN UNDER CURTAILED SAMPLING

Let us describe the expected sample size and duration for Simon's two-stage design under curtailed sampling. Let Y_1 be the smallest sample size needed for stage 1 of Simon's two-stage design under curtailed sampling. Then Y_1 is the SNB distributed sample size needed to observe either r_1 successes or $n_1 - r_1 + 1$ failures after following each patient for time t_2 .

Let Y_2 be the smallest sample size needed for stage 2 to reach a statistical endpoint under curtailed sampling. In studies terminating early, we have $Y_2=0$.

In trials not terminating early, the distribution of Y_2 is SNB(p_2, s, t), and Stage 2 sampling will stop as soon as we have observed either

$$s = r_2 - r_1$$

successes or else

$$t = n_1 + n_2 - y_1 + r_1 - r_2 + 1$$

failures for a given stage 1 curtailed sample size, y_1 .

In Simon's design, the number of successes observed in stage 1 counting towards rejecting the null hypothesis will be r_1 in trials continuing to Stage 2 under curtailed sampling.

The distribution of Y_2 can be expressed by conditioning on the values of the random variable Y_1 so

$$E(Y_2 | \text{continue to Stage 2}) =$$

$$\sum_{y_1=r_1}^{n_1} \Pr[Y_1 = y_1 | \text{continue to Stage 2}] \times E(Y_2 | y_1, \text{continue to Stage 2}),$$

Where

$$\Pr[Y_1 = y_1 | \text{continue to Stage 2}] \propto \binom{y_1-1}{r_1-1} (1-p_2)^{y_1}$$

follows a truncated negative binomial distribution for $r_1 \leq y_1 \leq n_1$.

The expected sample size for Simon's design under curtailed sampling is then

$$E(Y_1 + Y_2) = E(Y_1) + E(Y_2 | \text{continue to Stage 2}) \times (1 - \Pr[\text{early termination}]), \quad (C3)$$

where $\Pr[\text{early termination}] = \Pr[B < r_1]$ for $B \sim \text{Binomial}(n_1, p_2)$. In numerical studies, we found this expression yields the same values as reported by Chi and Chen.²¹

From (C3), we can also describe the expected duration of the trial. The expected duration of Simon's two-stage design under curtailed sampling is

$$E(Y_1 + Y_2)/\lambda + t_2 + t_1(1 - \Pr[\text{early termination}]).$$

In Simon's design, we have $r_2 > r_1$, so s and t given above will never take the value of 0. Under curtailed sampling, if the trial continues to stage 2, then additional patients are enrolled before a statistical endpoint is reached in Simon's design.

In curtailed sampling with the binomial design, the distribution of the sample size Y_2 is SNB with parameters $s=r_2$ and $t=n_2-r_2+1$. The expected duration of the binomial design is $E(Y_2)/\lambda + t_2$.

REFERENCES

1. Simon R. Optimal two-stage designs for. *Controlled Clinical Trials*. 1989;10(1):1–10. [PubMed: 2702835]
2. De With C.. Two-stage plans for the testing of binomial parameters. *Contemp Clin Trials*. 1983;4(3): 215–26.
3. Hanfelt JJ, Slack RS, Gehan EA. A modification of simon's optimal design for phase ii trials when the criterion is median sample size. *Control Clin Trials*. 1999;20(6):555–66. [PubMed: 10588296]
4. Jung SH, Carey M, Kim KM. Graphical search for two-stage designs for phase ii clinical trials. *Control Clin Trials*. 2001;22(4):367–72. [PubMed: 11514038]
5. Korn EL, Freidlin B. Conditional power calculations for clinical trials with historical controls. *Statistics in Medicine*. 2006;25(17):2922–31. [PubMed: 16479548]

6. Schill W, Wild P. Minmax designs for planning the second phase in a two-phase case-control study. *Statistics in Medicine*. 2006;25(10):1646–59. [PubMed: 16158403]
7. Thall PF, Simon R. Incorporating historical control data in planning phase ii clinical trials. *Statistics in Medicine*. 1990;9(3):215–28. [PubMed: 2188324]
8. Ye F, Shyr Y. Balanced two-stage designs for phase ii clinical trials. *Clin Trials*. 2007;4(5):514–24. [PubMed: 17942467]
9. DeMets DL. Current development in clinical trials: issues old and new. *Statistics in Medicine*. 2012;31(25):2944–54. [PubMed: 22736410]
10. Englert S. Flexible designs for single-arm phase ii trials in oncology. 2014.
11. Gehan EA, Schneiderman MA. Historical and methodological developments in clinical trials at the national cancer institute. *Statistics in Medicine*. 1990;9(8):871–80. [PubMed: 2218190]
12. Lachin JM. A review of methods for futility stopping based on conditional power. *Statistics in Medicine*. 2005;24(18):2747–64. [PubMed: 16134130]
13. Todd SA 25-year review of sequential methodology in clinical studies. *Statistics in Medicine*. 2007;26(2):237–52. [PubMed: 17139691]
14. Bryant J, Day R. Incorporating toxicity considerations into the design of two-stage phase ii clinical trials. *Biometrics*. 1995;51(4):1372–83. [PubMed: 8589229]
15. Kunz CU, Wason JMS, Kieser M. Two-stage phase ii oncology designs using short-term endpoints for early stopping. *Stat Methods Med Res*. 2017;26(4):1671–1683. [PubMed: 26037529]
16. Shih WJ, Ohman-Strickland PA, Lin Y. Analysis of pilot and early phase studies with small sample sizes. *Statistics in Medicine*. 2004;23(12):1827–42. [PubMed: 15195318]
17. Thall PF, Cheng SC. Treatment comparisons based on two-dimensional safety and efficacy alternatives in oncology trials. *Biometrics*. 1999;55(3):746–53. [PubMed: 11315002]
18. Thall PF, Cheng SC. Optimal two-stage designs for clinical trials based on safety and efficacy. *Statistics in Medicine*. 2001;20(7):1023–32. [PubMed: 11276033]
19. Ayanlowo AO, Redden DT. Stochastically curtailed phase ii clinical trials. *Statistics in Medicine*. 2007;26(7):1462–72. [PubMed: 16900560]
20. Bersimis S, Sachlas A, Papaioannou T. Flexible designs for phase ii comparative clinical trials involving two response variables. *Statistics in Medicine*. 2015;34(2):197–214. [PubMed: 25274584]
21. Chi Y, Chen CM. Curtailed two-stage designs in phase ii clinical trials. *Statistics in Medicine*. 2008;27(29):6175–89. [PubMed: 18816510]
22. Englert S, Kieser M. Adaptive designs for single-arm phase ii trials in oncology. *Pharm Stat*. 2012;11(3):241–9. [PubMed: 22411839]
23. Kunz CU, Kieser M. Curtailment in single-arm two-stage phase ii oncology trials. *Biochem J*. 2012;54(4):445–56.
24. Wu Y, Shih WJ. Approaches to handling data when a phase ii trial deviates from the pre-specified simon's two-stage design. *Statistics in Medicine*. 2008;27(29):6190–208. [PubMed: 18800338]
25. DeVeaux M, Kane MJ, Zelterman D. Clinical trial design using a stopped negative binomial distribution. *Statistics and its Interface*. 2018;11:699–707. [PubMed: 30655933]
26. Case LD, Morgan TM. Duration of accrual and follow-up for two-stage clinical trials. *Lifetime Data Anal*. 2001;7(1):21–37. [PubMed: 11280845]
27. Quoix E, Breton JL, Ducolone A, et al. First line chemotherapy with gemcitabine in advanced non-small cell lung cancer elderly patients: A randomized phase ii study of 3-week versus 4-week schedule. *Lung Cancer*. 2005;47(3):405–12. [PubMed: 15713524]

FIGURE 1.

Graphical schematic of notation for follow-up times, numbers of patients enrolled, outcome classifications, and their binomial probabilities. At the conclusion of stage 1 at follow-up time t_1 , there are X_1 successes, and we need a minimum of r_1 of these to continue to stage 2. At follow-up time t_2 , there are $X_{12} + X_2$ successes, and we need a minimum of r_2 of these to reject the null hypothesis

FIGURE 2.

Probability of rejecting the null hypothesis for the nested minmax/optimal design “A” in Table 1 for the lung cancer example. The locations of the null and alternative hypotheses are indicated

FIGURE 3.

Expected trial duration in a planned trial described in section 5 using the optimal nested design “A” (solid line), Simon optimal (dash), and Simon minimax (dots) using curtailed sampling for different values of the efficacy parameter p_2 and accrual rates λ .

TABLE 1

Details of selected nested designs with $n_1+n_2=36$ patients, maximum significance level of 0.1, futility parameter $p_1=0.8$, and null hypothesis $p_2=0.2$ suggested by the lung cancer example. Nested designs A to E are suggested in section 5. The expected sample size listed here for nested designs is the upper bound given at (2)

Design	n_1	n_2	r_1	r_2	α at $p_2=.2$	Expected Sample Size	Probability of Early Stopping	Power for $p_2=0.4$
A: Minimax/optimal	6	30	4	11	0.0834	33.0	.099	0.826
B	9	27	6	11	0.0847	33.7	.086	0.838
C	12	24	8	11	0.0858	34.3	.073	0.851
D	15	21	10	11	0.0866	34.7	.061	0.861
E: Balanced	18	18	12	11	0.0872	35.1	.051	0.870
X: Binomial	—	36	—	11	0.0889	36.0	0	0.910
Simon designs for $t_1=t_2$ and $p_1=p_2=.2$.								
Y: Simon optimal *	17	20	4	11	0.0948	26.0	.549	0.903
Z: Simon minimax	19	17	4	11	0.0861	28.3	.455	0.902

* Requires 37 patients.

TABLE 2

Expected sample size and trial duration to reach any statistical endpoint using curtailed sampling with the same parameter values and hypotheses as in Table 1. The null hypothesis H_0 is $p_2=.2$, and the alternative H_A is $p_2=.4$. The nested design uses parameters $t_1=2$ and $t_2=6$ months for follow-up times, and the Simon designs assume a 6 month follow-up. The expected trial duration in months is given for different rates λ of patient accrual per month

Design	n_1	n_2	r_1	r_2	Expected Curtailed		Expected Duration			
					Sample Size		Under H_0		Under H_A	
					Under H_0	Under H_A	$\lambda=2$	$\lambda=4$	$\lambda=2$	$\lambda=4$
A: minimax/optimal	6	30	4	11	29.3	24.6	22.1	14.7	19.7	13.6
B	9	27	6	11	29.9	25.1	22.4	15.0	20.0	13.8
C	12	24	8	11	30.4	25.5	22.8	15.2	20.3	13.9
D	15	21	10	11	30.8	25.9	23.1	15.3	20.6	14.1
E: balanced	18	18	12	11	31.2	26.2	23.3	15.5	20.8	14.2
Binomial	0	36	0	11	31.9	27.0	22.0	14.0	19.5	12.7
Simon designs for $t_1=t_2=6$ and $p_1=p_2=0.2$.										
Simon optimal	17	20	4	11	24.2	26.3	20.8	14.8	24.9	18.3
Simon minimax	19	17	4	11	26.3	26.6	22.4	15.8	25.2	18.5

TABLE 3

Comparison of Simon optimal, binomial, and nested optimal designs for different null and alternative hypotheses using curtailed sampling. The nested designs assume parameters $(.8, p)$ for p parameters of the specified null or alternative hypotheses. All designs have the same maximum sample size and nominal significance level .1, all optimized for the Simon design.

p Parameter Values									Expected Sample Size		Expected Duration			
H_0	H_A		n_1	n_2	r_1	r_2	α	Power	H_0	H_A	$\lambda=2$		$\lambda=4$	
0.05	0.25	Simon optimal	9	15	1	3	.0931	0.903	13.7	11.0	15.1	11.7	17.0	14.30
		Binomial	0	24	0	3	.0298	0.885	22.3	11.8	17.1	11.9	11.6	8.95
		Nested optimal	6	18	4	4	.0278	0.803	20.3	14.1	17.5	14.4	12.5	10.90
0.1	0.3	Simon optimal	12	23	2	6	.0977	0.901	18.5	18.5	17.3	12.7	20.7	16.1
		Binomial	0	35	0	6	.0552	0.935	32.3	19.9	22.1	15.9	14.1	11.0
		Nested optimal	6	29	4	7	.0514	0.846	29.2	20.9	22.0	17.9	14.7	12.6
0.2	0.4	Simon optimal	17	20	4	11	.0948	0.903	24.2	26.3	20.8	14.8	24.9	18.3
		Binomial	0	37	0	11	.0517	0.867	33.1	27.1	22.5	19.5	14.3	12.8
		Nested optimal	6	31	4	12	.0487	0.789	29.6	26.6	22.2	20.7	14.8	14.0
0.3	0.5	Simon optimal	22	24	8	18	.0974	0.905	27.3	34.2	21.6	14.8	28.7	20.2
		Binomial	0	46	0	18	.0681	0.908	40.6	35.7	26.3	23.8	16.2	14.9
		Nested optimal	6	40	4	19	.0643	0.825	36.2	34.0	25.5	24.4	16.5	15.9
0.4	0.6	Simon optimal	18	28	8	23	.0952	0.900	27.1	36.5	22.2	15.4	29.9	20.8
		Binomial	0	46	0	23	.0636	0.891	39.3	38.0	25.6	25.0	15.8	15.5
		Nested optimal	6	40	4	24	.0606	0.812	34.8	35.7	24.8	25.3	16.1	16.3
0.5	0.7	Simon optimal	21	24	12	27	.0963	0.902	25.3	36.7	20.6	14.3	29.9	20.8
		Binomial	0	45	0	27	.0676	0.901	37.2	38.2	24.6	25.1	15.3	15.6
		Nested optimal	6	39	4	28	.0650	0.824	32.8	35.8	23.8	25.3	15.6	16.4