STATISTICS IN MEDICINE
*Statist. Med.* 2002; **21**:1711–1726 (DOI: 10.1002/sim.1150)

# Early selection in a randomized phase II clinical trial‡

Seth M. Steinberg*,† and David J. Venzon

*Biostatistics and Data Management Section, Office of the Director, Center for Cancer Research, National Cancer Institute, National Institutes of Health, 6116 Executive Boulevard, Room 702, MSC 8325, Bethesda, MD 20892-8325, U.S.A.*

SUMMARY

A randomized phase II clinical trial design can be employed when one wishes to be able to select one of several similar therapies or variants of a therapy for inclusion in a subsequent, definitive, phase III trial. It is not necessary in this type of trial to formally identify a superior arm using the usual parameters and stringent criteria employed for hypothesis testing. Simon, Wittes and Ellenberg have described methods and sample sizes for selecting the superior arm from among $k$ possible arms. In this paper, we describe an approach based on statistical selection theory which allows one to potentially make a decision to end accrual to a randomized phase II trial at an interim point of the trial, and to select one of two arms as being worthy of further evaluation in a subsequent study. This method requires that an adequate gap in the number of responses between the two arms be observed at the interim point in order to limit the probability that the selected arm is actually inferior by more than an indifference amount. Published in 2002 by John Wiley & Sons, Ltd.

KEY WORDS: selection theory; randomized phase II trial; two-stage phase II trial

## 1. INTRODUCTION

It has been generally established over the past several decades that the process of oncologic drug development involves three phases of clinical trial [1]. Phase I trials are usually dose-seeking and dose-escalating studies conducted in order to identify the maximally tolerated dose (MTD) of an agent and thus an upper limit of a safe range in which to consider further administration of the agent. Phase II trials are usually intended to estimate the tumour response rate, in a specific patient population, of an agent at a dose determined by a phase I trial [1]. Often 75–90 per cent of the phase I MTD is selected as the dose to test in a phase II trial. Sample sizes of phase II trials are generally in the range of 30–50 for a single-arm study, and the trials are often conducted in two stages to permit an interim evaluation at which point insufficient responses to warrant continued interest in the agent will result in early termination

*Correspondence to: Seth M. Steinberg, Biostatistics and Data Management Section, Office of the Director, Center for Cancer Research, National Cancer Institute, National Institutes of Health, 6116 Executive Boulevard, Room 702, MSC 8325, Bethesda, MD 20892-8325, U.S.A.
†E-mail: ss118v@nih.gov
‡This article is a US Government work and is in the public domain in the U.S.A.

*Received May 2000*
Published in 2002 by John Wiley & Sons, Ltd. *Accepted September 2001*

of accrual [1, 2]. Phase II trials of combinations of agents also are commonly conducted. A phase III trial is a comparative randomized clinical trial which is usually the largest and most costly to conduct of the three phases of trials. In the case of a two-arm study, patients are usually randomized in equal proportions to receive one or the other treatment. The number of patients on the trial is determined in order to detect a difference in the endpoint of interest with a significance level of $\alpha$ and $(1-\beta)\times 100$ per cent power. Tables and nomograms which provide the number of patients needed to be enrolled into a phase III trial have been published [3–6], based on common endpoints, such as survival or response, and software to determine such sample sizes also exists [7–9].

In order to be selected for evaluation in a phase III trial, a therapeutic regimen will generally have undergone both phase I and II testing. In the case of a disease for which there is a standard therapy available, a new therapy with a well defined dose and route of administration that is found to have potentially better outcome characteristics in a phase II trial may be considered worthy of further evaluation in a phase III trial in comparison to the standard. Sometimes, however, the exact dose of an agent to evaluate in a phase III trial, or which of several promising therapies to advance first into a phase III trial, is unclear. For example, there is now an emerging area of development of vaccines for fighting cancer, and a recent successful early study may lead to many more such studies being conducted [10]. Questions about the route or mode of administration or dose schedule of the therapy may need to be addressed before phase III testing can begin.

The randomized phase II trial of Simon *et al.* is appropriate in cases like this [11]. Patients are enrolled centrally and then randomized to one of the two or more competing therapies. The arms can be conducted as if they were independent phase II trials, but the final results will be considered jointly for ranking the therapies or choosing one for further study. The purpose of this type of design is to simultaneously evaluate the efficacy of the therapies but not to formally compare them with the strict power and significance level of a phase III trial. This relaxation of requirements is permitted because one is merely making a rational choice of one arm and is not establishing statistically that the arm selected is superior. Clearly, the most appropriate setting for such an approach is one in which it would be desirable to select the better arm, but if, in fact, the arms are equivalent or the one which is chosen is slightly inferior to the ones which are not selected, the consequences of such a choice would not be too grave. In the cancer vaccine example above, a randomized phase II trial designed to decide whether an intradermal or intramuscular route of administration of a vaccine is to be used in future studies need not prove that one route is truly superior to the other. It is sufficient that the route chosen be at worst approximately equivalent (for example, only slightly inferior) to the one which was not chosen.

Sample sizes for the arms of a randomized phase II trial may be determined on the basis of standard phase II sample size considerations [1–3]. Another approach, discussed in the original presentation of the randomized phase II design, is to invoke statistical selection theory to design the size of a trial. Using this approach, one will simply select the arm with the best observed outcome to advance into later studies [12], and one determines the size of the study arms so as to guarantee a $p$ per cent probability of correctly selecting the superior arm if the superior arm has a true expected outcome that is at least a specified amount better than those of the inferior arms.

Throughout this paper, a binomial outcome, the response proportion, will be the only one considered, and the superior arm will be defined as the one for which the true response

probability exceeds that of the other (inferior) arm(s). As an example of selection theory in a two-arm trial, then, suppose the inferior response rate is believed to be 20 per cent and the superior is 35 per cent. In order to have 90 per cent probability of correctly selecting the superior arm, 29 patients per arm will be needed. The arm which turns out to have the higher observed response proportion after 29 patients per arm have been evaluated will be selected. If the response proportions are identical after 29 patients per arm, one arm is randomly chosen or secondary criteria could be invoked if appropriate.

By contrast, in order to have two-sided $\alpha = 0.05$ and 80 per cent power to detect a difference of 15 per cent (20 per cent versus 35 per cent) a study would require more than 130 patients per arm [7]. In selection theory, the concept of power does not have a direct analogue, but the probability of correctly choosing the superior arm is similar to the usual power. Randomized phase II trials, then, permit one to select an arm for subsequent comparative evaluation without requiring nearly so many patients as a rigorous definitive comparison.

As mentioned before, it is often desirable to employ an interim evaluation in order to justify continuation to the intended study size. In a phase II trial, interim evaluations are used in order to ensure that a therapy with a poor response rate early on is recognized and accrual is terminated. Two related designs developed by Thall *et al.* introduce the possibility of early termination of an experimental arm, although they apply more in a truly comparative setting. One is a two-stage design for first comparing all experimental arms to the control arm, and proceeding to a second stage only if an experimental arm is found to sufficiently exceed the control group's response rate [13]. Another proposal is a two-stage design for first choosing among $k$ experimental treatments and then possibly comparing the arm with the highest observed response rate to a control arm in a second stage if the highest experimental arm response rate is adequate [14].

In this paper, we propose an approach to performing one interim examination of the results from a randomized phase II two-arm trial. If the response proportions in the two arms are sufficiently different, accrual to the arm with fewer responses will stop early, and one may conclude with high probability that the arm selected is either superior to, or at worst more or less equivalent to, the other arm. Accrual to the selected arm will continue in such a case, up to the phase II trial size that it would attain in the absence of any alternatives. The notation and fundamental calculations for this approach are presented in the next section. Tables for the application of the method with a range of common parameter values appear in Section 3, followed by an example of a current trial incorporating this method and some extensions of the approach. In Section 6, we report some results based on the assumption of a constant odds ratio between response probabilities instead of a constant absolute difference.

## 2. EARLY SELECTION APPROACH

We will consider the setting of the two-arm randomized phase II trial and assume that the sample size for the two arms is determined either on the basis of statistical selection theory criteria or a standard phase II design approach [1–3]. Regardless of the size of the study, it will be considered desirable if, before reaching the intended accrual goal, one could select the arm with more responses for further evaluation if it is unlikely that it is actually inferior by more than a small amount. If the true probability of response in the superior arm is denoted $p_{\mathrm{H}}$ and the true probability of response in the inferior arm is $p_{\mathrm{L}}$, then the only error which

needs to be guarded against is selecting the arm which is truly inferior by more than $100\,g$ per cent, where $g = p_H - p_L$. Thus, $100\,g$ per cent is the size of the indifference region, that is, the maximum difference in response probabilities between the two arms for them to be considered essentially equivalent.

Let $n_1$ be the number of patients enrolled into one arm at some arbitrary point after the trial begins (say, $n_1 = n/2$, where $n$ is the total number of patients intended to be accrued per arm). Since 1:1 randomization is generally used between approximately equivalent arms, let $n_1$ also be the number of patients enrolled in the other arm at the same point. Define the number of observed responses on the superior arm as $r_H$ and the number of observed responses on the inferior arm as $r_L$. Let $d = |r_H - r_L|$ be the gap in the observed number of responses between the two arms after $2n_1$ patients have been enrolled. Thus $0 \leqslant d \leqslant n_1$. By selecting the arm with more responses when $d \geqslant d_E$, the smallest gap of interest, one may limit the probability of making an error of a particular magnitude as a function of the underlying true response probabilities. Thus, this design extends the selection approach described by Simon *et al.* [11] by requiring a gap of at least $d_E$ in the number of responses in order to permit a high probability of correct selection at an interim point in the trial.

If $r_H > r_L$, then the number of responses on the superior arm exceeds that on the inferior arm, and this is the desired outcome. There is no need to consider the probability of observing this set of outcomes since they are all correct. The problem is to limit the probability of making an error when $r_H - r_L \leqslant 0$, that is, when the responses are in the opposite direction from the underlying probabilities. The selection procedure permits one of the following three scenarios to occur: (i) selecting arm H when $r_H - r_L \geqslant d_E$; this would result in a correct decision; (ii) selecting arm L when $r_L - r_H \geqslant d_E$; then one has selected the inferior arm; (iii) making no decision when $|r_H - r_L| < d_E$; the trial then continues up to the intended accrual target.

Define $p_W$ as the probability of selecting the incorrect arm on the basis of interim responses. Then, under the assumption of independence between the numbers of responses of the two arms, for each value of $d$, $d = 0, \ldots, n_1$

$$\begin{aligned} p_W &= P(r_L - r_H \geqslant d \,|\, p_H, p_L) \\ &= P(r_L = d \,|\, p_L) P(r_H = 0 \,|\, p_H) \\ &\quad + P(r_L = d + 1 \,|\, p_L) P(r_H = 0; \text{or } r_H = 1 \,|\, p_H) \\ &\quad + \cdots \\ &\quad + P(r_L = n_1 \,|\, p_L) P(r_H = 0; \ldots; \text{ or } r_H = n_1 - d \,|\, p_H) \end{aligned}$$

If the response probabilities follow a binomial distribution, then

$$\begin{aligned} p_W = {} & b(d; p_L; n_1) B(0; p_H; n_1) + b(d+1; p_L; n_1) B(1; p_H; n_1) + \cdots \\ & + b(n_1; p_L; n_1) B(n_1 - d; p_H; n_1) \end{aligned}$$

where $b(k; p; n)$ is the standard binomial probability mass function for observing $k$ events of $n$ trials each with probability $p$, and $B(k; p; n)$ is the cumulative binomial probability of observing $\leqslant k$ out of $n$, each with probability $p$.

Thus, for $d = 0, \ldots, n_1$, and $p_H > p_L$, the probability of selecting the arm which is actually inferior is

$$p_W = \sum_{j=d}^{n_1} b(j; p_L; n_1) B(j - d; p_H; n_1) \tag{1}$$

It is easy to show that the value of $p_W$ is unchanged when $p_L$ is replaced by $1 - p_H$ and $p_H$ is replaced by $1 - p_L$ in the expression above. It follows from this symmetry that, for a fixed $g = p_H - p_L > 0$, numerical results for the interval $0 \leqslant p_L \leqslant (1 - g)/2$ can be reflected onto the interval $(1 - g)/2 \leqslant p_L \leqslant 1 - g$.

## 3. STRUCTURE OF TABLES

Applying formula (1), we used SAS release 6.12 (SAS Institute, Cary, NC) to generate a set of probabilities, $p_W$, of making an error in selecting an arm as a function of $d$, $p_L$, $p_H$ and $n_1$. Tables I and II show the minimal gaps, $d_E$, that must be equalled or exceeded after $n_1$ patients have been entered on each arm in order to be at least $100(1 - p_W)$ per cent correct at that point. Table I shows results when $g$ is 10 per cent (that is, it would be allowable for the 'wrong' arm to be selected, provided the true probabilities of response on the two arms differ by 10 per cent or less), and Table II is used if $g = 15$ per cent would be allowable. The tables are each ordered according to the underlying true response probabilities on the arms. Because of the symmetry in the value of $p_W$, only the entries up to $p_L = (1 - g)/2$ are needed, but higher values have been included for ease of reference.

In practice, of course, both $p_L$ and $p_H$ are unknown. For some study design methods [11], values of $p_L$ and $p_H$ must be assumed, but the only firm assumption we can make is that the width of the indifference region, $g$, can be fixed by clinical investigators. To be sure of an error rate no greater than $p_W$, one must employ the largest $d_E$ over all the values of $p_L$ within a reasonable range. In many cases, the response rates for standard therapies and from the phase I trials of the tested agents will restrict this range, but the largest $d_E$ over all $p_L$ will often have to be used.

This need not lead to extreme conservatism in the error rate. Figure 1 illustrates how, for some wide ranges of $p_L$ and corresponding $p_H$ values, $d_E$ remains the same for a given $n_1$ and $p_W$. For example, for $n_1 = 6$, as $p_L$ ranges between 0.18 and 0.67, and, with $g = 15$ per cent, $p_H$ ranges between 0.33 and 0.82 accordingly, the gap of 4 is needed to have 1 per cent maximal probability of selecting the inferior arm. Thus, as indicated by this figure, one can sometimes have great latitude in $p_H$ and $p_L$ with a fixed $d_E$, or one can simply use the largest $d_E$ shown for a fixed $n_1$ and $g$ and be certain that the error probability is maintained.

Tables I and II contain further conservatism of the sort seen in many statistical procedures with discrete outcomes and small sample sizes. For example, with $p_L = 0.4$, $p_H = 0.55$ and $n_1 = 6$, the gap of 3 is indicated to correspond to $p_W = 5$ per cent, and a gap of 2 corresponds to $p_W = 10$ per cent; however, the exact values in both cases, namely 2.4 per cent and 8.1 per cent, as determined by applying formula (1), are clearly less than these cut-off values.

Table III provides examples of the performance characteristics of the early selection design in comparison to a more standard single-stage randomized phase II design. Under each heading

Table I. Minimum observed differences between numbers of responses in two arms at an interim point in order to have less than $p_W$ probability of selecting inferior arm: $p_H - p_L = 0.10$.

| $p_W$ | $p_L$ | $p_H$ | $n_1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 5 | 6 | 8 | 10 | 12 | 15 | 20 |
| 0.5% | 0.05 | 0.15 | 3 | 3 | 3 | 3 | 3 | 3 | 4 |
| | 0.1 | 0.2 | 3 | 3 | 4 | 4 | 4 | 4 | 5 |
| | 0.2 | 0.3 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.3 | 0.4 | 4 | 5 | 5 | 5 | 6 | 6 | 7 |
| | 0.4 | 0.5 | 4 | 5 | 5 | 6 | 6 | 6 | 7 |
| | 0.5 | 0.6 | 4 | 5 | 5 | 6 | 6 | 6 | 7 |
| | 0.6 | 0.7 | 4 | 5 | 5 | 5 | 6 | 6 | 7 |
| | 0.7 | 0.8 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| 1% | 0.05 | 0.15 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.1 | 0.2 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| | 0.2 | 0.3 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| | 0.3 | 0.4 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.4 | 0.5 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.5 | 0.6 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.6 | 0.7 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.7 | 0.8 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| 5% | 0.05 | 0.15 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.1 | 0.2 | 2 | 2 | 2 | 3 | 3 | 3 | 3 |
| | 0.2 | 0.3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.3 | 0.4 | 3 | 3 | 3 | 3 | 4 | 4 | 4 |
| | 0.4 | 0.5 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| | 0.5 | 0.6 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| | 0.6 | 0.7 | 3 | 3 | 3 | 3 | 4 | 4 | 4 |
| | 0.7 | 0.8 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 10% | 0.05 | 0.15 | 2 | 2 | 2 | 2 | 2 | 1 | 1 |
| | 0.1 | 0.2 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.2 | 0.3 | 2 | 2 | 2 | 2 | 2 | 3 | 2 |
| | 0.3 | 0.4 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| | 0.4 | 0.5 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.5 | 0.6 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.6 | 0.7 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| | 0.7 | 0.8 | 2 | 2 | 2 | 2 | 2 | 3 | 2 |

of $n_1 = 5$ or $n_1 = 20$, the first column shows the probability of correct final selection with no provision for interim selection, assuming three sets of $p_L$, $p_H$ values and a full accrual of $n = 2n_1$ patients. As can be seen, the standard design provides approximately 65 to 95 per cent probability of correct selection when the two arms differ in the true probability of response by 10 to 15 per cent, for sample sizes of 10 and 40 subjects per arm. Under each sample size, the centre column indicates, for two extreme values of $p_W$, the difference between the probability in the first column and the probability of selection with an early evaluation, continuing to the end of planned accrual if not terminated early. The differences, ranging from 0 to 4.5 per cent depending on the parameter values, are substantially smaller than $p_W$, the error rate accepted at the interim evaluation. This occurs because many of the outcomes that result in incorrect

Table II. Minimum observed differences between numbers of responses in two arms at an interim point in order to have less than $p_W$ probability of selecting inferior arm: $p_H - p_L = 0.15$.

| $p_W$ | $p_L$ | $p_H$ | $n_1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 5 | 6 | 8 | 10 | 12 | 15 | 20 |
| 0.5% | 0.05 | 0.2 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.1 | 0.25 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| | 0.2 | 0.35 | 4 | 4 | 4 | 5 | 5 | 5 | 5 |
| | 0.3 | 0.45 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.4 | 0.55 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.5 | 0.65 | 4 | 4 | 5 | 5 | 5 | 6 | 6 |
| | 0.6 | 0.75 | 4 | 4 | 5 | 5 | 5 | 5 | 6 |
| | 0.7 | 0.85 | 4 | 4 | 4 | 4 | 4 | 5 | 5 |
| 1% | 0.05 | 0.2 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| | 0.1 | 0.25 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.2 | 0.35 | 3 | 4 | 4 | 4 | 4 | 4 | 4 |
| | 0.3 | 0.45 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| | 0.4 | 0.55 | 4 | 4 | 4 | 5 | 5 | 5 | 5 |
| | 0.5 | 0.65 | 4 | 4 | 4 | 5 | 5 | 5 | 5 |
| | 0.6 | 0.75 | 4 | 4 | 4 | 4 | 4 | 5 | 5 |
| | 0.7 | 0.85 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| 5% | 0.05 | 0.2 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |
| | 0.1 | 0.25 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.2 | 0.35 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.3 | 0.45 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.4 | 0.55 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.5 | 0.65 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.6 | 0.75 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.7 | 0.85 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| 10% | 0.05 | 0.2 | 1 | 1 | 1 | 1 | 1 | 1 | 1 |
| | 0.1 | 0.25 | 2 | 2 | 2 | 2 | 1 | 1 | 1 |
| | 0.2 | 0.35 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.3 | 0.45 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.4 | 0.55 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.5 | 0.65 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.6 | 0.75 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.7 | 0.85 | 2 | 2 | 2 | 2 | 2 | 2 | 1 |

early selection would end up with an incorrect final selection if the trial were to continue to full accrual. Finally, the rightmost column under each sample size heading presents the probability that a trial which uses the early selection approach will, in fact, terminate accrual early and correctly select the superior arm.

If the true difference between the arms is 10 or 15 per cent, there is only minimal probability of deriving benefit from early selection when $n_1 = 5$ and an extremely restrictive probability of being wrong is used ($p_W = 0.5$ per cent). On the other hand, for the higher values of these parameters, the probability of reducing the study size by correctly stopping early is between 10 and 90 per cent, and the decrease in the correct selection probability is small

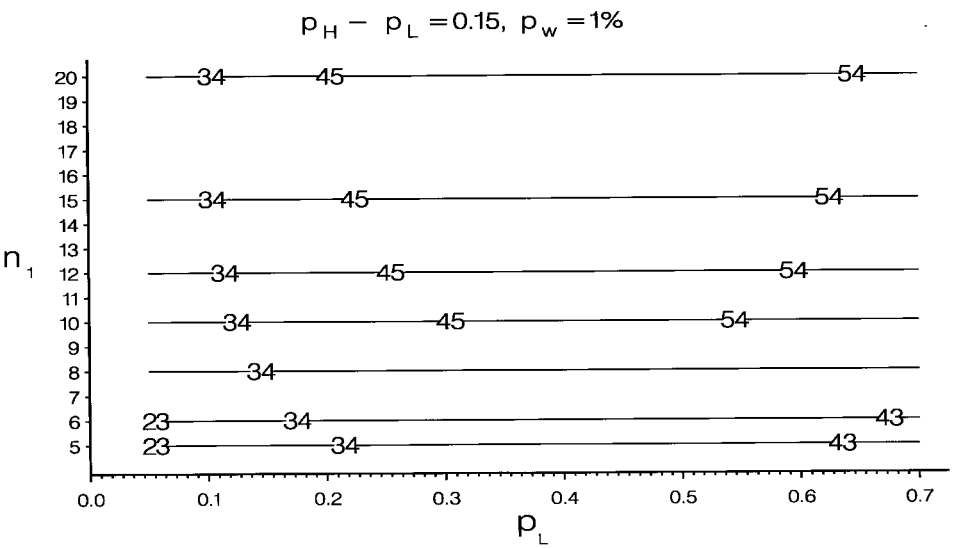


Figure 1. Minimum differences in observed responses needed for early selection with a maximum error probability $p_W = 1$ per cent for selected $n_1$ over a range of $p_L$, with $p_H - p_L = 0.15$.

to negligible depending on $p_W$. These results indicate that this procedure can be used in a beneficial manner with a wide range of trial design parameters.

## 4. EXAMPLE OF AN APPLICATION OF THE METHOD

A National Cancer Institute study of a DNA vaccine to be used along with Interleukin-2 (IL-2) was designed primarily to determine the clinical response rate of the combined therapy in patients with metastatic melanoma. However, a preliminary question was whether the vaccine is better administered intradermally or intramuscularly. To answer this first question, a small randomized phase II trial of the vaccine alone using a minimax design for each arm was approved [2]. Although this selection trial might have required as few as 12 patients per arm, the investigator wanted to be able to select one arm even more rapidly if this could be determined on the basis of early immune response findings. It was decided that 6 patients per arm would be used as the point at which this early evaluation would take place, since this was half the likely number that would be accrued per arm.

An error probability $p_W = 5$ per cent was the maximum which would be tolerated. In Table II, the gap of 3 is indicated as the largest minimal difference in the number of responses which must be observed in order to stop accrual early and select the arm with the higher number of observed responses, regardless of $p_H$ and $p_L$, provided that $p_H - p_L = 0.15$, to have less than 5 per cent probability of selecting the wrong arm at that point. Therefore, outcomes of 0/6 versus at least 3/6, 1/6 versus at least 4/6 and so forth will lead to early selection on the basis of immune responses. Otherwise, accrual will continue to the 12 patients accrual point.

Table III. Examples of comparisons of performance characteristics between randomized phase II designs with and without early selection.

| $p_L$ | $p_H$ | $p_W$ | $n_1 = 5,\ n = 10$ | | | $n_1 = 20,\ n = 40$ | | |
|---|---|---|---|---|---|---|---|---|
| | | | $P$(correct final selection: no interim selection) | Decrease with early selection* | $P$(early termination w/ correct selection) | $P$(correct final selection: no interim selection) | Decrease with early selection* | $P$(early termination w/ correct selection) |
| 0.05 | 0.15 | 0.005 | 0.763 | 0.000 | 0.021 | 0.936 | 0.000 | 0.205 |
| | | 0.10 | 0.763 | 0.000 | 0.133 | 0.936 | 0.037 | 0.791 |
| 0.4 | 0.5 | 0.005 | 0.672 | 0.000 | 0.023 | 0.816 | 0.000 | 0.074 |
| | | 0.10 | 0.672 | 0.000 | 0.098 | 0.816 | 0.010 | 0.439 |
| 0.7 | 0.8 | 0.005 | 0.694 | 0.000 | 0.011 | 0.849 | 0.000 | 0.098 |
| | | 0.10 | 0.694 | 0.002 | 0.228 | 0.849 | 0.021 | 0.574 |
| 0.05 | 0.20 | 0.005 | 0.843 | 0.000 | 0.046 | 0.983 | 0.001 | 0.589 |
| | | 0.10 | 0.843 | 0.014 | 0.575 | 0.983 | 0.024 | 0.896 |
| 0.40 | 0.55 | 0.005 | 0.748 | 0.000 | 0.033 | 0.911 | 0.000 | 0.213 |
| | | 0.10 | 0.748 | 0.004 | 0.321 | 0.911 | 0.025 | 0.686 |
| 0.70 | 0.85 | 0.005 | 0.787 | 0.000 | 0.015 | 0.948 | 0.000 | 0.281 |
| | | 0.10 | 0.787 | 0.002 | 0.277 | 0.948 | 0.045 | 0.833 |

* Difference between correct selection without interim selection and selection with interim evaluation, continuing to final if not stopped.

## 5. EXTENSIONS

### 5.1. Three or more arms

The early selection procedure can easily be extended to the case in which there are $q \geqslant 3$ arms. The simplest extension is to select the arm with the highest number of responses if that number is at least $d_E$ larger than the maximum number on the remaining $q-1$ arms. By a generalization of equation (1), the probability of making an incorrect selection of an arm can be obtained for the case where one arm has response probability $p_H$ and all the other arms have response probability $p_L$. In this case, each term in the summation is multiplied by the factor $(q-1)\{B(j-d; p_L; n_1)\}^{q-2}$ to account for the unselected inferior arms.

We determined the impact of the extension to three arms on the gaps that need to be equalled or exceeded in order to stop early with error probabilities ranging from 0.5 per cent to 10 per cent, as previously shown in Tables I and II. This revealed that the gaps for three arms generally remain the same as those for two arms, but some small gaps increase by one and some large gaps decrease by one. The probability of correct final selection is naturally lower with three arms than with two, even in the absence of interim selection; incorporating early selection appears to decrease this further by about $p_W$. In general, there is lower overall probability of early termination with three arms rather than two arms, and also a lower probability of early termination with correct selection. These decreases are not so large as to negate the benefit of early selection in this setting, however.

### 5.2. Combination with optimal phase II design

The early selection procedure can be applied to a randomized phase II trial with two arms in which each arm independently has the Simon optimal two-stage design [2]. At the end of the first stage, either or both arms are terminated if they fail to satisfy the criterion for that stage under the optimal design. Early selection takes effect only if both arms satisfy this criterion and one arm exceeds the other by the required gap or more. If one arm is selected and the other is terminated because of the early selection criterion, the remaining arm still must meet the optimal design criterion at the end of the second stage.

We can assess the effect of early selection on this design by determining how it changes the probability of the ultimate selection of the best-performing arm and the expected study size. An example is seen in Table IV. The two-stage optimal design has parameters $p_0 = 30$ per cent as an unacceptably low response probability and $p_1 = 60$ per cent as an acceptable probability, with error probability 0.10 both for mistakenly accepting an arm with true probability $p_0$ and for rejecting an arm with true $p_1$. A treatment is thus judged to be unacceptable if it has two or fewer responses out of the first eight (the first-stage criterion) or eight or fewer responses out of the total enrolment of 20. Early selection was based on $g = 0.15$, $p_W = 10$ per cent and the full range of $p_L$, so that from Table II, $d_E = 2$. Table IV displays some notable decreases in the expected total study size for higher response probabilities, at the cost of changes in selection probabilities of less than 0.02. From examples with other design parameters, it appears that the decreases in the final selection probabilities due to early selection are smaller when the study size is larger or when the true response probabilities are smaller, as expected from the decreased variance in the number of responses.

Table IV. Examples of selection probabilities under the two-stage optimal design without the early selection criterion (first row in each entry) and with early selection allowing an error probability $p_W = 10$ per cent (second row). Design parameters imply rejecting an arm for $\leqslant 2$ responses of 8 after the first stage or for $\leqslant 8$ responses of 20 after the second, final stage.

| True response probabilities for arm A, arm B | $P$(select A) | $P$(select B) | Expected study size |
|---|---|---|---|
| 0.3, 0.3 | 0.095 | 0.095 | 26.8 |
| | 0.094 | 0.094 | 26.4 |
| 0.5, 0.3 | 0.683 | 0.047 | 31.6 |
| | 0.677 | 0.048 | 30.1 |
| 0.5, 0.5 | 0.455 | 0.455 | 36.5 |
| | 0.452 | 0.452 | 33.5 |
| 0.7, 0.5 | 0.897 | 0.098 | 38.1 |
| | 0.881 | 0.113 | 32.9 |

### 5.3. *Optimal interim study size*

All the results presented up to this point have assumed that the study size at the interim analysis was fixed at $n_1$ in each arm, without any strict dependence on the intended full accrual of the trial, denoted by $n$. In principle, given the value of $n$, along with $g$, $p_W$ and a range for the $p_L$ values, one can determine the $n_1$ that meets any optimality criterion, simply by calculating the $n-1$ possible values and picking the best one.

One such criterion is minimizing the expected trial size. For each $n_1 < n$, this is calculated by finding first the largest $d_E$ for the given design parameters and then the probability $p_E$ of making an early selection of either arm using this $d_E$. This leads directly to the expected study size $n_E = p_E(n_1 + n) + (1 - p_E)2n$.

The practical application of this procedure may not be entirely as simple as this. First of all, $p_E$, and therefore $n_E$, are functions of $p_L$, and for certain parameter values the expected study size curves for different values of $n_1$ cross, so that no one optimal value can be determined for all $p_L$ within a designated range. This is seen in Figure 2 for $n = 25$, $g = 0.15$ and $p_W = 1$ per cent. The curves in the lower half of the figure show the optimal $n_1$ values of 8 for $0.30 < p_L < 0.50$, 16 for $0.20 < p_L < 0.30$ and 18 for $p_L$ near 0. Calculations with other parameter values show that such crossing of expected size curves becomes more of a complication as $n$ increases.

The jump from 8 to 16 in the optimal $n_1$ magnifies the difficulty of this case. It can be attributed to the discreteness in $d_E$ as a function of $n_1$, even for fixed $p_L$. The expected study size tends to increase or decrease smoothly as $n_1$ increases, but it can change substantially at those $n_1$ values where $d_E$ also increases by 1. Because of these jumps, multiple local minima may appear. An example of this is seen in the upper half of Figure 2, where jumps as $n_1$ increases from 3 to 4 and from 8 to 9 produce intermingled study size curves. This dependence on the discreteness of $d_E$ is of course more of a complication as $n$ decreases.

The narrow range of the vertical axis in Figure 2 shows that optimization would not yield a substantial reduction in the expected study size for the parameter values used there. This is a consequence of the small error probability $p_W$ and probability of early selection $p_E$.

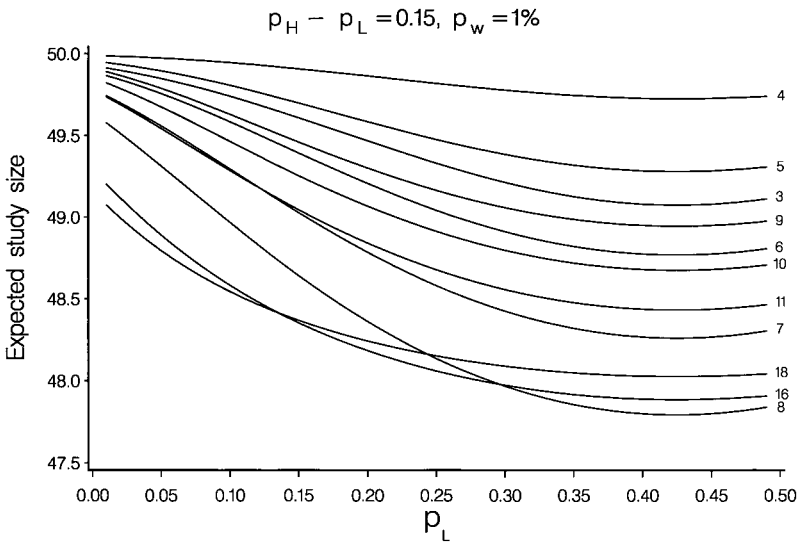


Figure 2. Expected study size for selected $n_1$ values (indicated to the right of the curves) derived from the probability of early selection between two arms with $n = 25$ each, $p_W = 1$ per cent and $p_H - p_L = 0.15$.

In addition, the families of expected study size curves are not always as ill-behaved as in Figure 2, but the characteristics displayed would at least have to be examined if minimizing the study size was important.

## 6. FORMULATION UNDER A CONSTANT ODDS RATIO MODEL

In our experience, clinical investigators are most familiar with comparing response rates in terms of their absolute difference, hence the parameter $g = p_H - p_L$ used above for the constant size of the indifference region between two arms. In some cases, however, it might be possible to define the indifference region in terms of the odds ratio $\gamma = p_H(1 - p_L)/\{p_L(1 - p_H)\} > 1$. In this section we will show that certain advantages are seen in these cases.

One way to apply this model is to construct tables comparable to Tables I and II, with $p_L$ varying over the unit interval and $p_H$ defined by the fixed odds ratio $\gamma$. Table V displays the gaps $d_E$ for this method, with the odds ratio $\gamma = 2$.

The variation over $p_L$ is narrower than in Tables I and II, especially for larger $n_1$ and larger $p_W$, and this is typical of other odds ratios we have tested. This is explained by the fact that the absolute difference $p_H - p_L$, as a function of $p_L$, is small for values of $p_L$ near 0, which requires larger gaps in this range than does a constant difference. The absolute difference increases to a maximum at $p_L = (\sqrt{\gamma} - 1)/(\gamma - 1)$ and then decreases again to a limit of 0 as $p_L$ approaches 1, keeping $d_E$ higher at the upper end of the range as well. As a consequence of this reduced variation in $d_E$, there is much less reason to be concerned about the unknown $p_L$ because the maximum $d_E$ over all $p_L$ will generally hold for a wider range of $p_L$ values.

Table V. Minimum observed differences between numbers of responses in two arms at an interim point in order to have less than $p_W$ probability of selecting inferior arm: $\gamma = 2$.

| $p_W$ | $p_L$ | $p_H$ | $n_1$ | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| | | | 5 | 6 | 8 | 10 | 12 | 15 | 20 |
| 0.5% | 0.05 | 0.10 | 3 | 3 | 3 | 3 | 4 | 4 | 4 |
| | 0.1 | 0.18 | 3 | 4 | 4 | 4 | 4 | 5 | 5 |
| | 0.2 | 0.33 | 4 | 4 | 4 | 5 | 5 | 5 | 5 |
| | 0.3 | 0.46 | 4 | 4 | 5 | 5 | 5 | 5 | 6 |
| | 0.4 | 0.57 | 4 | 4 | 5 | 5 | 5 | 5 | 6 |
| | 0.5 | 0.67 | 4 | 4 | 5 | 5 | 5 | 5 | 6 |
| | 0.6 | 0.75 | 4 | 4 | 5 | 5 | 5 | 5 | 6 |
| | 0.7 | 0.82 | 4 | 4 | 4 | 5 | 5 | 5 | 5 |
| | 0.8 | 0.89 | 3 | 4 | 4 | 4 | 4 | 5 | 5 |
| | 0.9 | 0.95 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| 1% | 0.05 | 0.10 | 3 | 3 | 3 | 3 | 3 | 4 | 4 |
| | 0.1 | 0.18 | 3 | 3 | 3 | 4 | 4 | 4 | 4 |
| | 0.2 | 0.33 | 3 | 4 | 4 | 4 | 4 | 5 | 5 |
| | 0.3 | 0.46 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| | 0.4 | 0.57 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| | 0.5 | 0.67 | 4 | 4 | 4 | 4 | 5 | 5 | 5 |
| | 0.6 | 0.75 | 4 | 4 | 4 | 4 | 4 | 5 | 5 |
| | 0.7 | 0.82 | 3 | 4 | 4 | 4 | 4 | 4 | 5 |
| | 0.8 | 0.89 | 3 | 3 | 4 | 4 | 4 | 4 | 4 |
| | 0.9 | 0.95 | 3 | 3 | 3 | 3 | 3 | 4 | 4 |
| 5% | 0.05 | 0.10 | 2 | 2 | 2 | 2 | 2 | 3 | 3 |
| | 0.1 | 0.18 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| | 0.2 | 0.33 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.3 | 0.46 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.4 | 0.57 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.5 | 0.67 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.6 | 0.75 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.7 | 0.82 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| | 0.8 | 0.89 | 2 | 2 | 3 | 3 | 3 | 3 | 3 |
| | 0.9 | 0.95 | 2 | 2 | 2 | 2 | 2 | 3 | 3 |
| 10% | 0.05 | 0.10 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.1 | 0.18 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.2 | 0.33 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.3 | 0.46 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.4 | 0.57 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.5 | 0.67 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.6 | 0.75 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.7 | 0.82 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.8 | 0.89 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |
| | 0.9 | 0.95 | 2 | 2 | 2 | 2 | 2 | 2 | 2 |

The odds ratio formulation also permits a potentially more useful criterion for early selection. It results from evaluating the error probability $p_W$ conditional on the total number of observed responses at the interim analysis, which eliminates the dependence on $p_L$. Following Cox [15], we let $T = r_L + r_H$, and the error probability for any

Table VI. Minimum differences between numbers of observed responses in two arms at an interim point, $n_1 = 12$, in order to have less than $p_W$ probability of selecting inferior arm, given that the odds ratio of $p_H$ to $p_L$ is $\gamma = 2$.

| $p_W$ | Larger of the response numbers on two arms | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 |
| 0.5% | | | 5 | 5 | 6 | 6 | 6 | 6 | 5 | 5 |
| 1% | | 4 | 5 | 5 | 5 | 5 | 5 | 5 | 5 | 4 |
| 5% | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 | 3 |
| 10% | 3 | 3 | 3 | 2 | 2 | 2 | 3 | 3 | 3 | 3 |

$d = \mathrm{mod}(T, 2), \ldots, T$ is

$$P(r_L - r_H = d \,|\, T, \gamma) = \frac{\binom{n_1}{(T+d)/2}\binom{n_1}{(T-d)/2}\gamma^{(T-d)/2}}{\sum_{u=\max(0,\,T-n_1)}^{\min(n_1,\,T)} \binom{n_1}{T-u}\binom{n_1}{u}\gamma^u} \tag{2}$$

As $T$ takes on each value between 0 and $2n_1$, we determine the smallest $d_E$ such that the sum of expression (2) over all $d \geqslant d_E$ is no greater than $p_W$. This $d_E$ does not exist for all $T$, since we know, as an extreme example, that all the conditional probability is concentrated at $d = 0$ for $T = 0$ and $T = 2n_1$.

The gaps conditional on $T$ that allow early selection are more concisely expressed as a function of the larger of the two numbers of responses in the two arms. Typical results are then shown in Table VI, using $n_1 = 12$ and $\gamma = 2$. For some parameter values, $d_E$ is the same for any observed outcome, and in these cases the early selection criterion has the same form as the one using the maximum over $p_L$ in Table I or Table II. More often, however, $d_E$ truly does depend on the observed interim outcome. We note that with uniform weighting over the interval $0.05 \leqslant p_L \leqslant 0.95$, the average absolute difference $p_H - p_L$ is approximately 0.12 for $\gamma = 2$, and indeed, the entries of Table VI alternate between the maximum $d_E$ entries in the $n_1 = 12$ columns of Table I, with $p_H - p_L$ fixed at 0.10, and Table II, with $p_H - p_L = 0.15$.

This conditional probability criterion shares the usual disadvantage of conservatism in its error probabilities. The number of terms in the denominator of expression (2) is small for some values of $T$, and this discreteness often results in an exact probability that is substantially less than $p_W$. This may limit the value of this approach when $n_1$ is small or the true response rate is believed to be near the extremes of the unit interval.

The generalization of expression (2) to three or more arms is not as straightforward as the generalization of (1) is, since additional combinatorial coefficients with dependencies on the sum $T$ must be factored into the numerator and the denominator. The practical results are qualitatively similar, however.

## 7. DISCUSSION

Statistical selection theory is one of the approaches which can be used to determine the sample size for the selection of the arm with greater observed responses from a randomized phase II clinical trial. By extending the concept to require a gap of $d_E$ responses rather than one response, small probabilities of making the error of selecting a truly inferior arm can be

achieved with surprisingly few patients. If we consider the randomized phase II setting in which it is a quite reasonable objective to select either arm for subsequent study provided adequate numbers of patients have been evaluated, this extension seems logical and may be beneficial. This is particularly true whenever a decision can and should be made on the basis of small to moderate numbers of patients, the consequences of selecting the wrong arm are not great, and the indifference region is of moderate size, say 10–15 per cent at most. Of course, this also assumes that other factors such as toxicity do not materially affect the decision.

This study demonstrates that $d_E$ can be determined on the basis of hypothesized $p_H$ and $p_L$, or that the gap can be bounded and thus would not require any more information to specify than the sample size, easily $n/2$, the size of the indifference region and the probability one is willing to accept of an error in selecting an inferior arm at an interim point.

Information about the true underlying response rates $p_H$ and $p_L$ may be very sparse, and we have described two methods for eliminating the dependence of the early selection design on these parameters. Conditioning on the observed total number of responses is more robust to parameter misspecification than is maximizing over a range of $p_L$ values, but it requires an indifference region defined by an odds ratio. We see this as generally limiting its applicability, but in cases where clinical investigators are sufficiently knowledgeable or where suitable approximations can be made, it may be the preferred method. On the other hand, many conditions are known to have either very good or very poor response rates to all current therapies, and in such cases the maximum $d_E$ over a restricted range in $p_L$ may produce a higher probability of early selection of one arm.

The inclusion of the early selection procedure in the design phase of a study has implications for the estimation of the confidence interval around the difference in the observed response proportions. This confidence interval should be adjusted for interim evaluations performed, possibly requiring extension of current methods [16, 17].

## ACKNOWLEDGEMENTS

The authors would like to thank Dr Richard Simon for his assistance in the development of this project, Dr Paul Albert for his helpful advice and critique of the manuscript, a reviewer for suggesting several of the topics included in the paper, and David Liewehr for technical assistance.

## REFERENCES

1. Leventhal BG, Wittes RE. *Research Methods in Clinical Oncology*. Raven: New York, 1988.
2. Simon R. Optimal two-stage designs for phase II clinical trials. *Controlled Clinical Trials* 1989; **10**:1–10.
3. Simon RM. Clinical Trials in Cancer. In *Cancer*: *Principles and Practice of Oncology*, 5th edn, Devita VT Jr, Hellman S, Rosenberg SA (eds). Lippencott-Raven: New York, 1997; 513–528.
4. George SL, Desu MM. Planning the size and duration of a clinical trial studying the time to some critical event. *Journal of Chronic Diseases* 1974; **27**:15–24.
5. Rubinstein LV, Gail MH, Santner TJ. Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *Journal of Chronic Diseases* 1981; **34**:469–479.
6. Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 1982; **38**:163–170.
7. Dupont WD, Plummer WD Jr. Power and sample size calculations: a review and computer program. *Controlled Clinical Trials* 1990; **11**:116–128.
8. Bernstein D, Lagakos SW. Sample size and power determination for stratified clinical trials. *Journal of Statistical Computing and Simulation* 1978; **8**:65–73.
9. Halpern J, Brown BW Jr. Designing clinical trials with arbitrary specification of survival functions and for the log rank or generalized Wilcoxon test. *Controlled Clinical Trials* 1987; **8**:177–189.

10. Rosenberg SA, Yang JC, Schwartzentruber DJ *et al.* Immunologic and therapeutic evaluation of a synthetic peptide vaccine for the treatment of patients with metastatic melanoma. *Nature Medicine* 1998; **4**:321–327.
11. Simon RM, Wittes RE, Ellenberg S. Randomized phase II clinical trials. *Cancer Treatment Reports* 1985; **69**:1375–1381.
12. Gibbons JD, Olkin I, Sobel M. *Selecting and Ordering Populations*: *A New Statistical Methodology*. Wiley: New York, 1977.
13. Thall PF, Simon R, Ellenberg S. Two-stage selection and testing designs for comparative clinical trials. *Biometrika* 1988; **75**:303–310.
14. Thall PF, Simon R, Ellenberg S. A two-stage design for choosing among several experimental treatments and a control in clinical trials. *Biometrics* 1989; **45**:537–547.
15. Cox DR. *The Analysis of Binary Data*. Chapman and Hall: London, 1970; 48–49.
16. Atkinson EN, Brown BW. Confidence limits for probability of response in multistage phase II clinical trials. *Biometrics* 1985; **41**:741–744.
17. Duffy DE, Santner TJ. Confidence intervals for a binomial parameter based on multistage tests. *Biometrics* 1987; **43**:81–93.