A Two-Stage Design for Choosing among Several Experimental Treatments and a Control in Clinical Trials

Peter F. Thall; Richard Simon; Susan S. Ellenberg

Biometrics, Vol. 45, No. 2 (Jun., 1989), 537-547.

Stable URL:
http://links.jstor.org/sici?sici=0006-341X%28198906%2945%3A2%3C537%3AATDFCA%3E2.0.CO%3B2-E

Biometrics is currently published by International Biometric Society.

Your use of the JSTOR archive indicates your acceptance of JSTOR's Terms and Conditions of Use, available at http://www.jstor.org/about/terms.html. JSTOR's Terms and Conditions of Use provides, in part, that unless you have obtained prior permission, you may not download an entire issue of a journal or multiple copies of articles, and you may use content in the JSTOR archive only for your personal, non-commercial use.

Please contact the publisher regarding any further use of this work. Publisher contact information may be obtained at http://www.jstor.org/journals/ibs.html.

Each copy of any part of a JSTOR transmission must contain the same copyright notice that appears on the screen or printed page of such transmission.

For more information on JSTOR contact jstor-info@umich.edu.

©2003 JSTOR

http://www.jstor.org/
Fri Nov 21 10:00:43 2003

BIOMETRICS 45, 537–547
June 1989

A Two-Stage Design for Choosing Among Several Experimental Treatments and a Control in Clinical Trials

Peter F. Thall

Statistics/Computer & Information Systems Department,
George Washington University, Washington, D.C. 20052, U.S.A.

Richard Simon

Biometric Research Branch, Cancer Therapy Evaluation Program, DCT, NCI, NIH,
6130 Executive Boulevard, Rockville, Maryland 20892, U.S.A.

and

Susan S. Ellenberg

Biostatistics Research Branch, AIDS Program, NIAID, NIH,
6003 Executive Boulevard, Rockville, Maryland 20892, U.S.A.

SUMMARY

In clinical trials where several experimental treatments are of interest, the goal may be viewed as identification of the best of these and comparison of that treatment to a standard control therapy. However, it is undesirable to commit patients to a large-scale comparative trial of a new regimen without evidence that its therapeutic success rate is acceptably high. We propose a two-stage design in which patients are first randomized among the experimental treatments, and the single treatment having the highest observed success rate is identified. If this highest rate falls below a fixed cutoff then the trial is terminated. Otherwise, the "best" new treatment is compared to the control at a second stage. Locally optimal values of the cutoff and the stage-1 and stage-2 sample sizes are derived by minimizing expected total sample size. The design has both high power and high probability of terminating early when no experimental treatment is superior to the control. Numerical results for implementing the design are presented, and comparison to Dunnett's (1984, in *Design of Experiments: Ranking and Selection*, T. J. Santner and A. C. Tamhane (eds), 47–66; New York: Marcel Dekker) optimal one-stage procedure is made.

1. Introduction

In clinical research there are often several experimental treatments of interest for evaluation. The objective of a program of clinical trials in this setting may be viewed as identification of a single most effective new treatment and determination of the merits of this treatment relative to a standard "control" therapy. This is motivated by circumstances where there are insufficiently many patients to conduct a large-scale trial comparing all of the experimental treatments to each other and to the control. In most situations, preliminary evidence concerning the levels of effectiveness and relative merits of the experimental treatments is imperfect. Such evidence typically arises from uncontrolled pilot studies, often from several different sources. In practice, judgments concerning which experimental treatment to compare to the standard therapy in a statistically controlled clinical trial often are made in an informal manner. These judgments are based on the pilot data and possibly also prior

Key words: Clinical trial; Early termination; Least favorable configuration; Sample size; Two-stage design.

information concerning the effectiveness of the standard therapy. Unfortunately, actual differences among the experimental treatments may be confounded with systematic differences among the pilot studies. This severely limits the usefulness of such data in comparing and evaluating the experimental treatments for the purpose of selecting one for subsequent comparison to the control.

Still, it is generally accepted by clinicians that comparison of a single experimental treatment to an established therapy requires a randomized clinical trial. Given some specified Type I error rate, the sample size is determined to achieve a specified power at some simple alternative to a composite null hypothesis. However, when this is carried out subsequent to the type of informal selection procedure described above, the nominal power figure may be misleadingly high. This is due to the fact that the preliminary selection process has been ignored in the power computation.

In this paper we build upon the conventional approach described above by including the preliminary selection process in the trial design. Our motivation arises from the desire to obtain practical designs that avoid the confounding described earlier while controlling the overall power of the procedure.

Formally, we consider the problem of first choosing one of K experimental treatments and then determining whether it is superior to a control treatment C. We shall deal with the binomial setting where each patient's treatment response may be characterized as success or failure, and denote by θ_k the success probability of the kth experimental treatment, with $k = 0$ corresponding to C. For notational convenience we assume $\theta_1 \leqslant \theta_2 \leqslant \cdots \leqslant \theta_K$ and denote $\boldsymbol{\theta} = (\theta_0, \theta_1, \ldots, \theta_K)$.

The design proposed here is carried out in two stages. At the first stage, patients are randomized equally among the K experimental treatments. The highest observed success rate is compared to a fixed cutoff λ, which is based on a prespecified estimate θ_0^* of θ_0. If this rate is greater than or equal to λ, then the corresponding "best" treatment is selected for a subsequent randomized two-arm comparison to the control. Otherwise, the trial is terminated. The two stages are thus essentially a selection followed by a comparative test, with possible early termination prior to the second stage.

It is important to emphasize that we do not treat the case where it is possible to make a decision immediately after each individual patient's treatment response. In such a circumstance, an adaptive allocation or selection procedure, such as a "play the winner" rule (cf. Kulkarni and Jennison, 1986) could be considered. Such an approach may be difficult to apply, however, in clinical trials where patient response is not observable for weeks or months following study entry. When multiple clinical centers are involved it would be complicated logistically to implement such an adaptive procedure, even if patient response were immediate. The general methods proposed here, however, could be optimized further through the use of more fully sequential methods.

The design is defined formally in Section 2. A generalized definition of power to accommodate the particular structure of the design is introduced in Section 3. Section 4 describes derivation of the design parameters. Numerical results including specific designs for various values of K, θ_0^*, and overall power are presented in Section 5, and these results are discussed in Section 6.

2. The Design

Choose constants δ_1 and δ_2 so that $\theta_0 + \delta_1$ represents a marginal improvement over θ_0, while $\theta_0 + \delta_2$ is a clinically significant improvement. Our proposed design is defined as follows:

Stage 1. Randomize Kn_1 patients equally among the K experimental treatments. If the most successful treatment has a proportion of successes greater than or equal to λ, then it progresses to stage 2; otherwise, the trial is terminated.

Stage 2. Randomize an additional $2n_2$ patients equally to C and to the treatment ν selected at stage 1. Perform a one-sided test of $\theta_\nu = \theta_0$ versus $\theta_\nu > \theta_0$.

For $0 < \delta_1 < \delta_2 < 1 - \theta_0^*$, the stage-1 power requirement is that if some $\theta_k \geqslant \theta_0^* + \delta_2$ and no θ_k is between $\theta_0^* + \delta_1$ and $\theta_0^* + \delta_2$, then a treatment with $\theta_k \geqslant \theta_0^* + \delta_2$ will progress to stage 2 with probability greater than or equal to β_1. The stage-2 test is carried out with size α and power greater than or equal to β_2 for $\theta_0 = \theta_0^*$ and $\theta_\nu = \theta_0^* + \delta_2$. In the event that it is concluded $\theta_\nu > \theta_0$, we shall say the experimental treatment ν has been "chosen."

Our approach is especially motivated by circumstances where most experimental therapies will not provide a clinically significant improvement over the standard therapy. Unfortunately, this situation is common in cancer clinical trials. An initial estimate θ_0^* of θ_0 based on prior clinical experience with the standard treatment typically is available. This enables us to derive a cutoff λ so that the stage-1 selection procedure allows the possibility of early termination. We include this possibility in the design because it is undesirable to carry out a full-scale trial of a new treatment that is not reasonably promising compared to the standard therapy. For the designs derived here, it will be seen that the numerical value of λ is approximately $\theta_0^* + .10$; hence, the stage-1 selection procedure depends in an essential way on θ_0^*.

In any case, a selected experimental treatment must demonstrate a statistically significant improvement over the standard in a full-scale randomized trial before it may be declared a clinical improvement. Thus, the second stage must include a control arm, even though the initial estimate θ_0^* of θ_0 was used in the earlier selection process. The stage-2 test may be carried out via a Fisher exact procedure or using normal approximations to the distributions of the sample proportions. Performing the stage-2 test at level α is equivalent to requiring a conditional size α for the overall procedure, given that a treatment has progressed to stage 2. Traditionally, it is desired to ensure a size no larger than the nominal level α regardless of what the actual value of θ_0^* may be. Use of the conditional size here guarantees this, since otherwise the value of λ and hence θ_0^* necessarily would be involved in the computation of α. The actual overall size of each of the designs derived here is thus below the nominal size, and so the procedure is conservative in this regard.

Although the two stages are carried out independently, they are connected by use of θ_0^* both in determining the value of λ and in the power computation at stage 2 (cf. Lachin, 1981; Casagrande, Pike, and Smith, 1978). The other connection between the two stages is that, under a least favorable configuration of θ to be defined below, β_1 and β_2 must satisfy the constraint $\beta = \beta_1\beta_2$, where β is the overall generalized power. We shall derive the values of β_1 and β_2 along with λ, n_1, and n_2 in the derivation of the design for specified β. Analogous derivations also are carried out with both β and β_2 specified.

In order to implement the design, the values of n_1, n_2, and λ must be determined. The total sample size N is random, taking on the value Kn_1 or $Kn_1 + 2n_2$ as the trial either terminates early or continues to a second stage. We shall determine the values of the design parameters that minimize the expected sample size $E(N)$ for given $\theta_0 = \theta_0^*$, subject to the size and power constraints described above.

3. The Least Favorable Configuration

The power constraint described in the previous section deals with configurations of the success parameters such that at least one experimental treatment is "acceptable"; i.e., some $\theta_k \geqslant \theta_0^* + \delta_2$, and none of the experimental treatments have success parameters in the band $(\theta_0^* + \delta_1, \theta_0^* + \delta_2)$. For any such configuration, we require that the probability of choosing an acceptable treatment must be at least β. Among this set of configurations, however, the probability of choosing an acceptable treatment varies considerably. For example, this probability will be great when all the experimental treatments have success probabilities

close to unity. We define a configuration of the set to be "least favorable" if it minimizes the probability of an acceptable choice. The general concept of least favorable configuration (LFC) is discussed in Gibbons, Olkin, and Sobel (1977). The following theorem characterizes the LFC for our problem.

Theorem 1 Under the above conditions, a LFC is $(\theta_0^*, \theta_0^* + \delta_1, \ldots, \theta_0^* + \delta_1, \theta_0^* + \delta_2)$, where the entry $\theta_0^* + \delta_1$ is repeated $K - 1$ times.

A proof is given in the Appendix. The theorem says that a LFC has only one experimental treatment with success probability greater than or equal to $\theta_0^* + \delta_2$, and its success probability is at the lower boundary of this region. The other $K - 1$ experimental treatments have success probabilities at the upper boundary of the region less than or equal to $\theta_0^* + \delta_1$.

If we require the power to be at least β under the LFC, where $\theta_0 = \theta_0^*$, $\theta_1 = \cdots = \theta_{K-1} = \theta_0^* + \delta_1$, and $\theta_K = \theta_0^* + \delta_2$, then we will have automatically assured that the power is at least β for any configuration in which $\theta_0 = \theta_0^*$, $\theta_k \notin (\theta_0^* + \delta_1, \theta_0^* + \delta_2)$ for any $k = 1, \ldots, K$, and at least one experimental treatment has $\theta_k \geqslant \theta_0^* + \delta_2$. Under the LFC $\boldsymbol{\theta}^* = (\theta_0^*, \theta_0^* + \delta_1, \ldots, \theta_0^* + \delta_1, \theta_0^* + \delta_2)$, the power, or probability of an acceptable choice, may be written as the product $\beta = \beta_1\beta_2$, where β_1 and β_2 are the respective probabilities that (i) the best treatment is selected at the end of the first stage and $\hat{\theta}_K > \lambda$ at that point, and (ii) the hypothesis $\theta_0 = \theta_K$ is rejected in favor of $\theta_0 < \theta_K$ at the end of the second stage.

Let X_k denote the number of successes with treatment k at stage 1, $\hat{\theta}_k = X_k/n_1$ the corresponding observed success rate, $b(n, \theta; \cdot)$ the binomial probability mass function with parameters n and θ, and $B(n, \theta; \cdot)$ the corresponding cumulative distribution function. The explicit formula for stage-1 power under the LFC is

$$\beta_1 = \Pr_*\Big(\hat{\theta}_K \geqslant \max_{1 \leqslant k < K} \hat{\theta}_k, \text{ experimental treatment } K \text{ wins the randomization in case of a tie, and } \hat{\theta}_K \geqslant \lambda\Big)$$

$$= \sum_{x=n_0}^{n_1} \sum_{j=0}^{K-1} (j+1)^{-1} C(K-1, j) b(n_1, \theta_0^* + \delta_2; x) \times b(n_1, \theta_0^* + \delta_1; x)^j B(n_1, \theta_0^* + \delta_1; x-1)^{K-1-j}, \quad (1)$$

where $C(K-1, j)$ is the combinatoric term for choosing j out of $K - 1$ treatments and n_0 is the smallest integer not smaller than λn_1. Since $\hat{\theta}_K = X_K/n_1 \geqslant \lambda$, the first summation is over $X_K \geqslant n_0$. The index j accounts for the number of suboptimal experimental treatments, i.e., those having success probability $\theta_0^* + \delta_1$, which are tied with the best treatment K for the highest observed response rate after stage 1. The factor $(j + 1)^{-1}$ is the probability that treatment K wins the randomization among the tied experimental treatments. The stage-2 power β_2 is the usual value associated with the test of $\theta_0 = \theta_K$ versus $\theta_0 < \theta_K$ at alternative $\theta_0 = \theta_0^*$, $\theta_K = \theta_0^* + \delta_2$.

The expected total sample size is $\mathrm{E}(N) = Kn_1 + \pi 2n_2$, where π is the probability of continuing to the second stage. Under the simple null hypothesis $\theta_0 = \theta_1 = \cdots = \theta_K = \theta_0^*$, the continuation probability is

$$\pi_0 = 1 - \{B(n_1, \theta_0^*; n_0 - 1)\}^K. \quad (2)$$

Under the LFC the continuation probability is

$$\pi_* = 1 - \{B(n_1, \theta_0^* + \delta_1; n_0 - 1)\}^{K-1} B(n_1, \theta_0^* + \delta_2; n_0 - 1). \quad (3)$$

For the designs to be developed here, the power will be calculated under the LFC. One may question how well such designs perform under other configurations of $\boldsymbol{\theta}$. If more than one θ_k exceeds $\theta_0^* + \delta_2$ and the others are less than or equal to $\theta_0^* + \delta_1$, then the probability of choosing one of the acceptable treatments will exceed β. However, these designs are aimed at choosing one acceptable experimental treatment. The probability of choosing the best of several acceptable experimental treatments cannot be guaranteed large, especially if the success probability of the second best is close to that of the best.

Of course, choosing a regimen with response rate in the interval $(\theta_0^* + \delta_1, \theta_0^* + \delta_2)$ still represents identification of an improved treatment. If a treatment has a success rate in the lower portion of the interval, then the probability of selecting it over a superior treatment (with rate greater than or equal to $\theta_0^* + \delta_2$) will be low. If a regimen has a success rate in the upper portion of the interval, then choosing it instead of one having rate greater than or equal to $\theta_0^* + \delta_2$ is not a serious error. The following theorem gives a bound on the power of the procedure in the more general situation where, under the assumption that at least one experimental treatment has success rate no less than $\theta_0^* + \delta_2$, there may be treatments with θ_k inside this interval.

Theorem 2 If $\theta_K \geqslant \theta_0^* + \delta_2$, then the probability of choosing a treatment k with $\theta_k \geqslant \theta_0^* + \delta_1$ is bounded below by $\beta_1\beta_2(\delta_1)$, where β_1 is as before and $\beta_2(\delta_1)$ is the stage-2 power at $\theta_0 = \theta_0^*$ and $\theta_k = \theta_0^* + \delta_1$.

Thus, if one is willing to consider a treatment with success rate greater than or equal to $\theta_0^* + \delta_1$ to be acceptable, then the procedure still has power no smaller than $\beta_1\beta_2(\delta_1)$. A proof of Theorem 2 is given in the Appendix.

4. Derivation of Designs

4.1 *Expected Total Sample Size*

A natural criterion for choosing n_1, n_2, and cutoff λ is minimization of the expected sample size $\mathrm{E}(N)$. The expected sample size, however, depends on the success probabilities. If the expected sample size is minimized under the simple null hypothesis $\theta_0 = \theta_1 = \cdots = \theta_K = \theta_0^*$, then λ is very large, so that one is unlikely to proceed to the second stage. Since the constraint $\beta_1\beta_2 \geqslant \beta$ must be satisfied, the stage-1 power β_1 approaches β and β_2 approaches 1. Hence, minimization of the expected sample size $\mathrm{E}_0(N)$ under the null hypothesis results in a design with an impractically large stage-2 sample size. If the expected sample size $\mathrm{E}_*(N)$ under the LFC $\boldsymbol{\theta}^*$ is minimized, then λ is forced to be negative and a second stage becomes certain. This design has poor performance under the null hypothesis. A compromise that avoids such extreme designs is to minimize the weighted average $\omega\mathrm{E}_0(N) + (1 - \omega)\mathrm{E}_*(N)$. We have adopted this approach. The designs presented in our tables are based on $\omega = \frac{1}{2}$, although other weights could be used. This point is discussed briefly in Section 6.

4.2 *Designs for Fixed β*

We first minimize $\mathrm{E}(N)$ by fixing θ_0^* and β, with the probabilities β_1 and β_2 allowed to vary subject to the constraint $\beta_1\beta_2 = \beta$. Given n_1 and λ, β_1 is computed from (1). Note that since the lower limit of the first summation in (1) is the smallest integer not less than λn_1, β_1 is a step function in λ. The stage-2 sample size n_2 is then computed via the method of Casagrande et al. [1978, expression (8)]. Since n_1 and λ also determine the continuation probabilities given in expressions (2) and (3), $\mathrm{E}(N)$ may then be computed. The

minimization was carried out via a global search with n_1 incremented in an outer loop and $E(N)$ minimized in λ for each n_1.

4.3 *Designs for Both β and β_2 Fixed*

In some circumstances one might wish to specify the power of the stage-2 test, in addition to the overall power. To accommodate this additional constraint, the preceding approach may be modified by first fixing β and β_2 at the desired values, with the step function β_1 required to satisfy $\beta_1 \geqslant \beta/\beta_2$. The global search is carried out by incrementing n_1 for a small initial value of λ until admissible β_1 is obtained, then minimizing $E(N)$ in λ for this n_1. The minimum is obtained by iterating this procedure. In this case β_2 determines n_1. If λ were fixed, n_1 would simply be the smallest value yielding admissible β_1.

5. Results

We shall illustrate the properties of the proposed two-stage design for the cases $\theta_0^* =$ .2, .4, .6; $\delta_1 = .05$; $\delta_2 = .20$; $\alpha = .05$; $K = 2, 3, 4$; and β in the range .70 to .80. Although power figures of .80 to .90 are common for Phase III trials of a single experimental treatment, the β here represents both selection of the appropriate experimental treatment and subsequent demonstration that it is significantly better than the control.

For the case with α and β specified, numerical results are presented in Table 1. A range of values of λ is given for each design, since $E(N)$ is a step function of λ. The designs in Table 1 appear quite practical and have some good properties. First, the selection stage is

Table 1
Designs for fixed power β and size $\alpha = .05$, $\delta_1 = .05$, $\delta_2 = .20$

θ_0^*	K	n_1	n_2	λ	$E_0(N)$	$E(N)$	N_{max}	$1-\pi_0$	β_1	β_2	β
.2	2	28	89	.290–.320	86.6	150.3	234	.8280	.8041	.8704	.70
		32	97	.285–.310	97.0	168.4	258	.8284	.8381	.8949	.75
		40	99	.280–.300	113.0	190.0	278	.8326	.8869	.9021	.80
.2	3	31	98	.295–.320	133.5	202.4	289	.7925	.7785	.8992	.70
		38	102	.290–.315	149.7	227.0	318	.8245	.8231	.9112	.75
		45	111	.290–.310	167.9	256.6	357	.8518	.8580	.9324	.80
.2	4	34	105	.295–.320	183.1	256.6	346	.7740	.7631	.9173	.70
		41	109	.295–.315	205.9	288.0	382	.8073	.8082	.9280	.75
		48	120	.295–.310	231.1	325.9	432	.8362	.8441	.9478	.80
.4	2	33	98	.485–.515	110.5	177.5	262	.7725	.8250	.8485	.70
		39	108	.490–.510	119.8	198.6	294	.8063	.8515	.8808	.75
		47	119	.490–.510	131.3	224.1	332	.8430	.8800	.9091	.80
.4	3	38	111	.505–.525	162.0	239.3	336	.7828	.7876	.8888	.70
		44	123	.505–.520	177.9	268.3	378	.8123	.8182	.9166	.75
		51	129	.495–.505	205.7	302.5	411	.7952	.8612	.9290	.80
.4	4	40	120	.505–.525	223.7	304.1	400	.7341	.7681	.9113	.70
		48	128	.505–.520	248.0	340.3	448	.7795	.8099	.9260	.75
		56	142	.505–.515	275.6	384.5	508	.8172	.8437	.9482	.80
.6	2	26	86	.695–.730	91.3	149.3	224	.7717	.8148	.8591	.70
		32	90	.690–.715	102.8	166.6	244	.7822	.8612	.8709	.75
		39	99	.695–.715	111.1	187.3	276	.8313	.8875	.9014	.80
.6	3	30	94	.705–.730	138.1	200.6	278	.7436	.7885	.8878	.70
		37	100	.705–.725	151.1	223.9	311	.7987	.8288	.9049	.75
		43	107	.700–.720	170.6	251.7	343	.8055	.8661	.9237	.80
.6	4	34	99	.710–.735	187.6	253.7	334	.7379	.7763	.9017	.70
		41	105	.710–.730	207.3	283.8	374	.7919	.8178	.9171	.75
		47	113	.705–.720	233.7	319.2	414	.7976	.8545	.9362	.80

relatively small; the values of n_1 are in the range commonly employed in clinical oncology for uncontrolled Phase II studies. Of course, smaller values of $\delta_2 - \delta_1$ would require larger stage-1 sample sizes, but the value $\delta_2 - \delta_1 = .15$ employed here is applicable in many situations.

The stage-2 sample sizes and powers are also consistent with those employed in many Phase III trials. The stage-2 power β_2 is greater than the stage-1 power β_1 (the LFC probability of selecting the one acceptable treatment K and concluding a second stage), and this condition maintains small expected sample size when the null hypothesis is true. The probability $1 - \pi_0$ of stopping after the first stage when the null hypothesis is true is very high, ranging from .73 to .85 for the designs in Table 1. This is a very desirable characteristic, since we would not wish to allocate $2n_2$ additional patients to a second stage in such circumstances. Although the power and the probability of proceeding to the second stage are calculated for $\theta = \theta_0^*$, the Type I error rate α is never exceeded for any θ_0. It is also interesting to note that the derived values of the cutoff criterion λ used for determining whether to proceed to the second stage all are approximately $\theta_0^* + .10$, for these δ_1 and δ_2. This result, as well as the observed relationship between β_1 and β_2, could be used to heuristically generate nearly optimal designs for other values of θ_0^* in the range .2 to .6, given $\delta_1 = .05$ and $\delta_2 = .20$.

Some statisticians might take the view that the second stage should be designed with a power of .80, say, regardless of the size of the first stage and regardless of the chance that the first stage might have failed to select an appropriate regimen. Table 2 shows results for

Table 2

Designs for fixed power β, stage-2 test power β_2, and size $\alpha = .05$, $\delta_1 = .05$, $\delta_2 = .20$, $\theta_0^* = .20$, subject to $\beta_1 \geq \beta/\beta_2$

K	n_1	n_2	λ	$E_0(N)$	$E(N)$	N_{max}	$1 - \pi_0$	β_1	β_2	β
2	37	74	.275–.295	103.2	158.4	222	.8017	.8800	.80	.70
	34	84	.295–.320	88.1	152.7	236	.8797	.8235	.85	.70
	30	98	.305–.330	83.2	155.5	256	.8816	.7923	.90	.70
	51	74	.255–.270	136.6	191.9	250	.7644	.9380	.80	.75
	35	84	.260–.285	115.3	173.4	238	.7298	.8848	.85	.75
	32	98	.285–.310	97.6	170.2	260	.8284	.8381	.90	.75
	54	84	.260–.275	141.9	207.6	276	.7978	.9423	.85	.80
	43	91	.260–.275	131.6	197.0	268	.7473	.9162	.875	.80
	38	98	.265–.285	120.8	192.5	272	.7711	.8921	.90	.80
3	45	74	.270–.285	174.7	226.9	283	.7203	.8787	.80	.70
	36	84	.280–.305	148.8	208.2	276	.7563	.8266	.85	.70
	31	98	.295–.320	133.6	202.7	289	.7925	.7785	.90	.70
	67	74	.270–.280	227.4	287.1	349	.8202	.9389	.80	.75
	48	84	.275–.290	182.7	245.3	312	.7691	.8874	.85	.75
	41	98	.295–.315	152.0	228.8	319	.8517	.8340	.90	.75
	68	84	.265–.275	238.0	304.2	372	.7974	.9420	.85	.80
	58	91	.280–.290	203.8	277.4	356	.8345	.9151	.875	.80
	51	98	.275–.290	191.6	267.9	349	.8026	.8953	.90	.80
4	52	74	.270–.285	250.5	301.9	356	.7111	.8763	.80	.70
	43	84	.280–.300	216.0	275.5	340	.7374	.8318	.85	.70
	37	98	.300–.320	185.4	257.7	344	.8086	.7785	.90	.70
	75	74	.270–.280	332.4	389.5	448	.7796	.9377	.80	.75
	55	84	.275–.290	261.8	323.6	388	.7509	.8855	.85	.75
	46	98	.285–.300	227.9	300.8	380	.7756	.8436	.90	.75
	79	84	.280–.290	337.6	410.0	484	.8711	.9417	.85	.80
	67	71	.285–.295	292.4	368.9	450	.8647	.9148	.875	.80
	59	98	.290–.305	263.0	344.8	432	.8618	.8891	.90	.80

the case with β_2 specified in advance along with α and β, for $\theta_0^* = .20$. In order to achieve the same overall β in Table 2 compared to Table 1, larger values of β_1 tend to be required in Table 2.

Table 3 compares the two-stage designs developed here with some single-stage designs previously studied by Dunnett (1984), for $\theta_0^* = .2$. Although Dunnett's designs provide more information about experimental treatments not chosen as best, their performance under the null hypothesis is much inferior to the two-stage designs. While the maximum sample size of this procedure exceeds that of Dunnett for $K = 2$, when $K = 3$ or $K = 4$ even our maximum sample size is about the same as or less than Dunnett's. The comparison of these two classes of designs is imperfect, however, because the two designs do not provide the same power for alternatives other than the LFC and because Dunnett employed normal approximations that are known to overestimate power in other circumstances (Casagrande et al., 1978). Analogous comparisons for $\theta_0^* = .4$ and $\theta_0^* = .6$ yielded very similar results.

Table 3
Comparison of required sample sizes for Dunnett's single-stage and proposed two-stage designs at $\alpha = .05$, $\theta_0 = .20$, $\delta_1 = .05$, $\delta_2 = .20$

		Dunnett	Two-stage		
	β	N_t	$E_0(N)$	$E(N)$	N_{max}
$K = 2$	.70	189	86.6	150.3	234
	.75	213	97.0	168.4	258
	.80	240	113.0	190.0	278
$K = 3$	.70	284	133.5	202.4	289
	.75	316	149.7	227.0	318
	.80	356	167.9	256.6	357
$K = 4$	.70	385	183.1	256.6	346
	.75	425	205.9	288.0	382
	.80	475	231.1	325.9	432

6. Discussion

The numerical results for the proposed two-stage designs are encouraging. A key to the potential usefulness of this design is the very high probability of early termination at the end of stage 1 when the global null hypothesis is true. A less quantitative, but equally important feature, is the potential improvement in practical trial design resulting from selecting which experimental treatment to test based on a randomized first stage rather than on multiple uncontrolled pilot studies. There are advantages to both approaches, but increased use of randomized pilot studies would be beneficial (Simon, Wittes, and Ellenberg, 1985).

This design is obviously not applicable to all clinical situations. First, we have assumed that the outcome is binary and observed soon enough after initiation of treatment to allow the interim stage-1 decision. The binary outcome assumption could be relaxed, but the need to make decisions based on data observed relatively soon after treatment limits the applicability of this approach when the outcome is survival. In some clinical trials, there is interest in determining the effectiveness of each experimental treatment relative to the control. Our design is not appropriate in such situations. However, one must recognize that the usual designs for evaluating K experimental treatments and a control do not address the issue of selecting among regimens that are found to be improvements.

Since the numerical values of the stage-1 cutoff λ are all approximately $\theta_0^* + .10$, the stage-1 power depends on the initial value θ_0^* of θ_0. If θ_0^* is above or below the true value

of θ_0, the actual value of β_1 will be below or above its nominal value, respectively. Thus, if the value of θ_0^* is taken to be too high, this has the consequence of setting a stage-1 cutoff that is above the supposed value, hence increasing the probability of missing an improvement over the control at stage 1. In contrast, if $\theta_0^* < \theta_0$, there is an increased risk of carrying out a stage-2 comparison when ideally one would wish to terminate the trial early. At stage 2, the effect of incorrectly specifying θ_0^* is the same as in a conventional test based on binomial data, since θ_0^* appears as a nuisance parameter in β_2.

The minimization of $E_\theta(N)$ was carried out under the prior placing equal weights on $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}^*$. More elaborate priors could be used. One possible approach would be to take θ_0 to be beta on an interval centered around θ_0^*, and $\theta_1, \ldots, \theta_K$ beta on a larger interval containing both θ_0^* and $\theta_0^* + \delta_2$, with all $K + 1$ parameters mutually independent. Alternatively, one could attempt to elicit priors without specifying a functional form. However, prior distributions for treatment effects generally are quite variable among clinicians. Hence, tabulation of results for a limited set of priors did not seem useful. If one regards our objective function from a Bayesian viewpoint, placing equal weights on the null and LFC values of $\boldsymbol{\theta}$ does reflect a rather simple and easily understood prior. However, our aim in selecting an objective function was simply to obtain practical designs with appealing properties.

Some related designs should be mentioned. Whitehead (1986) considered the problem of selecting among experimental treatments and then comparing the selected treatment to a randomized control. His design, however, does not permit early termination after the first stage. His approach is Bayesian and requires that a prior distribution be specified for the success probabilities of the experimental treatments. Whitehead assumes that there are a large number of experimental treatments and seeks to quantify the tradeoff between more treatments and more patients per treatment arm for a fixed total number of patients. Whitehead's formulation is motivated by a particular type of clinical situation in which patient numbers are limited relative to the large number of treatments available for study. It is an interesting approach, but would not apply to many clinical situations for which the design presented here would be appropriate.

Ellenberg and Eisenberger (1985) and Lan, Simon, and Halperin (1982) have developed designs that permit early termination of a trial when interim results for an experimental treatment are not sufficiently promising. Although studied only for the case of a single experimental treatment, their approaches could be generalized for interim comparison of each experimental treatment to the control. Thall, Simon, and Ellenberg (1988) have proposed a design, similar to that studied here, which includes a control arm at the first stage. This results in a larger overall stage-1 sample size, larger $E_0(N)$, and smaller expected sample size under the alternative, since stage-1 data are utilized in the final comparison of the selected experimental treatment to the control. However, such an approach may be less appropriate under the circumstances described earlier, where experimental treatments are frequently found to offer no improvement over the standard therapy. Furthermore, the concept of requiring evidence of promise for a regimen before starting a randomized comparison to a control is appealing to many clinical investigators. Consequently, the two-stage design proposed here may be more acceptable in many situations than designs that employ a control from the start. In either case, incorporation of the selection process into the randomized trial should serve to reduce potential selection bias and confounding while controlling overall error probabilities.

Acknowledgements

The authors wish to thank an associate editor and two referees for their constructive comments and suggestions.

Résumé

L'évaluation de plusieurs nouveaux traitements par des essais thérapeutiques peut être considéré comme l'identification du meilleur d'entre eux et la comparaison de ce dernier à un traitement de référence. Néanmoins, il n'est pas souhaitable de débuter un essai comparatif sans s'être assuré d'une efficacité suffisante des nouvelles thérapeutiques. Nous proposons un plan d'expérience en deux phases. Dans un premier temps, les patients sont randomisés entre les nouveaux traitements. Celui permettant d'obtenir le taux de succés le plus élevé est retenu. Si ce taux est inférieur à un seuil fixé au départ, l'essai est terminé. Sinon, au cours d'une deuxième étape, le "meilleur" des nouveaux traitements est comparé au traitement de référence. Les valeurs optimales du seuil et des effectifs nécessaires aux phases un et deux sont obtenues en minimisant l'effectif attendu de l'échantillon total. Ce plan d'expérience a une puissance élevée et la probabilité de conclure précocement est importante quand aucun des nouveaux traitements n'est supérieur au traitement de référence. Des exemples d'utilisations de ce plan d'expérience sont présentés et comparés à la procédure optimale en une étape de Dunnett (1984, dans *Design of Experiments: Ranking and Selection*, T. J. Santner et A. C. Tamhane (eds), 47–66; New York: Marcel Dekker).

References

Casagrande, J. T., Pike, M. C., and Smith, P. G. (1978). An improved approximate formula for calculating sample sizes for comparing to binomial distributions. *Biometrics* **34,** 483–486.

Dunnett, C. W. (1984). Selection of the best treatment in comparison to a control with an application to a medical trial. In *Design of Experiments: Ranking and Selection*, T. J. Santner and A. C. Tamhane (eds), 47–66. New York: Marcel Dekker.

Ellenberg, S. S. and Eisenberger, M. A. (1985). An efficient design for Phase III studies of combination chemotherapies (with Discussion). *Cancer Treatment Reports* **69,** 1147–1154.

Gibbons, J. D., Olkin, I., and Sobel, M. (1977). *Selecting and Ordering Populations: A New Statistical Methodology.* New York: Wiley.

Kulkarni, R. V. and Jennison, C. (1986). Optimal properties of the Bechhofer–Kulkarni Bernoulli selection procedure. *Annals of Statistics* **14,** 298–314.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* **2,** 93–113.

Lan, K. K. G., Simon, R., and Halperin, M. (1982). Stochastically curtailed tests in long-term clinical trials. *Communications in Statistics—Sequential Analysis* **1,** 207–214.

Simon, R., Wittes, R. E., and Ellenberg, S. S. (1985). Randomized Phase II clinical trials. *Cancer Treatment Reports* **69,** 1375–1381.

Thall, P. F., Simon, R., and Ellenberg, S. S. (1988). Two-stage selection and testing designs for comparative clinical trials. *Biometrika* **75,** 303–310.

Whitehead, J. (1986). Sample sizes for Phase II and Phase III clinical trials: An integrated approach. *Statistics in Medicine* **5,** 459–464.

Received February 1987; *revised December* 1987, *and August and November* 1988.

Appendix

A.1 Proof of Theorem 1

Write $\theta_0^* = \theta_0$ for convenience. Let $A = \{k \geqslant 1: \theta_k \geqslant \theta_0 + \delta_2\}$, the set of indices of acceptable experimental treatments, so that $A \cup A^c = \{1, \ldots, K\}$. For given $\boldsymbol{\theta}$,

$$\begin{aligned}\Pr_{\boldsymbol{\theta}}(\text{Acceptable choice}) &= \Pr_{\boldsymbol{\theta}}(\text{Reject at stage 2} \mid \text{Acceptable selection at stage 1}) \\ &\quad \times \Pr_{\boldsymbol{\theta}}(\text{Acceptable selection at stage 1}) \\ &= \beta_2(\boldsymbol{\theta})\beta_1(\boldsymbol{\theta}).\end{aligned}$$

Note that β_2 depends on $\boldsymbol{\theta}$ only through θ_0 and the selected treatment's success rate θ_ν. Denote the cardinality of A by $|A| = m$, so that $|A^c| = K - m$. For simplicity assume $A^c \neq \varnothing$, i.e., $K - m \geqslant 1$, since the case $A^c = \varnothing$ is subsumed trivially by the argument to follow. We have assumed $K \in A$, so

$1 \leqslant m \leqslant K - 1$ and consequently the LFC must satisfy the general condition

$$\theta_1 \leqslant \cdots \leqslant \theta_{K-m} \leqslant \theta_0 + \delta_1, \qquad \theta_0 + \delta_2 \leqslant \theta_{K-m+1} \leqslant \cdots \leqslant \theta_K. \tag{B}$$

Recall that $\beta_1(\boldsymbol{\theta})$ is the probability under $\boldsymbol{\theta}$ that $\max\{X_k: k \in A\} \geqslant \max\{n_0, \max\{X_j: j \in A^c\}\}$ and an element of A wins the randomization if there is a tie. But $\max\{X_k: k \in A\}$ and $\max\{X_j: j \in A^c\}$ are stochastically increasing in each element of $\{\theta_k: k \in A\}$ and each element of $\{\theta_j: j \in A^c\}$, respectively. [A random variable W is *stochastically increasing* in parameter γ if $\Pr_\gamma(W > c)$ is a nondecreasing function of γ for all values of c.] Thus, $\beta_1(\boldsymbol{\theta})$ must be minimized among $\boldsymbol{\theta}$ satisfying condition B at $\theta_1 = \cdots = \theta_{K-m} = \theta_0 + \delta_1$ and $\theta_{K-m+1} = \cdots = \theta_K = \theta_0 + \delta_2$. It thus remains to show that minimization in m occurs at $m = 1$, since $\beta_2(\boldsymbol{\theta})$ already is minimized when θ_ν achieves its lower bound $\theta_0 + \delta_2$, for acceptable treatment ν.

Suppose $m > 1$. Let $U_1, \ldots, U_K$ be independent random variables identically distributed uniformly on $[0, \varepsilon]$, where ε is a fixed number between 0 and 1. Defining $Y_k = X_k + U_k$, we may express

$$\beta_1(\boldsymbol{\theta}) = \Pr_{\boldsymbol{\theta}}[\max\{Y_k: k \in A\} > \max\{n_0, \max\{Y_k: k \in A^c\}\}],$$

since replacing each X_k by Y_k does not alter any inequalities among the X_k's but acts as a fair randomization device in the case of a tie. If one element is removed from A and placed into A^c, this has the effect of stochastically decreasing $\max\{Y_k: k \in A\}$ and stochastically increasing $\max\{Y_k: k \in A^c\}$, thus decreasing β_1. The minimum is achieved by iterating until only $m = 1$ element remains in A.

A.2 Proof of Theorem 2

Let $A_0 = \{k \geqslant 1: \theta_k \leqslant \theta_0 + \delta_1\}$, $A_1 = \{k \geqslant 1: \theta_0 + \delta_1 \leqslant \theta_k \leqslant \theta_0 + \delta_2\}$, and $A_2 = A$ as given in the proof of Theorem 1. We take $\{A_0, A_1, A_2\}$ to be a partition of $\{1, \ldots, K\}$, with any index k having θ_k on a boundary placed either way without loss of generality. Define $M_i = \max\{Y_k: k \in A_i\}$, $i = 0, 1, 2$, where Y_k is the version of X_k as given in the previous proof. An acceptable selection occurs at stage 1 if $\max\{M_1, M_2\} > \max\{M_0, n_0\}$. As before, M_i is stochastically increasing in θ_k for each $k \in A_i$, $i = 0, 1, 2$; hence, the probability of an acceptable choice is bounded below by taking $\theta_k = \theta_0 + \delta_2$ for all $k \in A_2$, and $\theta_k = \theta_0 + \delta_1$ for all $k \in A_0 \cup A_1$, then switching elements from A_2 to $A_0 \cup A_1$, until $A_2 = \{K\}$. Since $\beta_2 > \beta_2(\delta_1)$, the desired result follows.