The Annals of Statistics
1993, Vol. 21, No. 1, 464–483

A SEQUENTIAL CLINICAL TRIAL FOR COMPARING THREE TREATMENTS[1]

BY D. SIEGMUND

Stanford University

A model for sequential clinical trials involving three treatments is proposed and various error probabilities are defined. A specific procedure involving a concatenation of repeated significance tests is studied in detail. The first phase of testing proceeds until at least one treatment is eliminated; and then a second phase is used to compare the remaining treatments. Analytic approximations to the operating characteristics of the test are derived and compared with the results of a Monte Carlo experiment.

1. Introduction. This paper is concerned with sequential clinical trials for comparing three treatments in the simplest possible situation that patient responses are instantaneous, independent, normally distributed random variables with a common known variance and possibly different mean values. The building blocks for our test are modified repeated significance tests discussed in detail by Siegmund [(1985), Chapters IV and V]. It is certainly possible and perhaps interesting to make a similar study with the O'Brien–Fleming (1979) stopping rule as the basic building block.

The procedure can be described roughly as follows. We begin sampling from each of the three treatments and use the repeated significance test discussed by Siegmund [(1980), (1985), Chapter V]. If that test indicates the existence of some treatment effect, the least promising treatment is discarded and testing continues to compare the remaining two treatments. Our goal will be to indicate how various error probabilities and expected sample sizes depend on the test parameters and to get some idea how much less power and larger expected sample size this procedure has than a two treatment trial in which the worst treatment has been eliminated a priori.

The evaluation of our procedure is much more complex than for the comparison of two treatments. In general terms the interesting situations are as follows.

(i) All three treatments are equally effective. An ideal test would not eliminate any treatment. To do so would be to make a Type I-1 error.

(ii) Two treatments are equally effective while the third is inferior. Failure to eliminate the inferior treatment would be a Type II-1 error. Elimination of one of the two superior treatments would be a Type I-2 error.

Received March 1991; revised February 1992.

[1]Research supported by AFOSR Grant 89-0355, NSA Grant MDA 904-89-H-2040 and NSF Grant DMS-89-05874.

AMS 1991 *subject classifications*. Primary 62L10; secondary 62N10.

Key words and phrases. Sequential test, boundary crossing probability, repeated significance test.

(iii) The treatments can be clearly ranked in order of preference. Our goal is to select the best treatment, and failure to do so would be a Type II-2 error.

(iv) One treatment is the best and the other two are about equally effective. Our goal is the same as in (iii), but attaining it is more difficult.

The evaluation of expected sample sizes is also more complicated, since there is first the expected number of triples of observations needed to eliminate at least one treatment and then the expected number of additional pairs of observations required to reach a final conclusion.

The paper is organized as follows. Section 2 is concerned with describing more precisely the test and the measures of its quality we shall consider. Section 3 contains the results of a Monte Carlo study. Section 4 gives analytic approximations aimed at gaining insight into some of the results in Section 3. Open problems are discussed in Section 5.

We begin with a brief description of repeated significance tests to compare two treatments. Assume $Z_1, Z_2, \ldots$ are independent and normally distributed with mean μ and unit variance. The Z's are differences in response of paired subjects, one of whom receives treatment 1 and the other of whom receives treatment 2. Let $S_n = \sum_{i=1}^{n} Z_i$, $n = 1, 2, \ldots$. Given $b \geq c > 0$ and positive integers $m_0 < m$, a repeated significance test of the hypothesis that $\mu = 0$ stops sampling at $\min(T, m)$, where

$$T = \inf\{n: n \geq m_0, |S_n| > bn^{1/2}\},$$

and rejects the hypothesis of no treatment effect if either $T < m$ or $T \geq m$ and $|S_m| > cm^{1/2}$. The probability of rejecting the null hypothesis is

$$\begin{aligned} &P_\mu\{T < m\} + P_\mu\{T \geq m, |S_m| > cm^{1/2}\} \\ &\quad = P_\mu\{|S_m| > cm^{1/2}\} + P_\mu\{T < m, |S_m| < cm^{1/2}\}. \end{aligned} \tag{1}$$

Siegmund [(1985), Chapter IV] gives approximations for these probabilities and for the expected sample size $E_\mu[\min(T, m)]$. In particular it is shown that by a judicious choice of the test parameters b, c and m_0 one can have a test whose power function is very close to that of a fixed sample test of sample size m while taking on average only a small fraction of m observations when $|\mu|$ is large. This entails taking c small compared to b, and then the first term on the right-hand side of (1) is much larger than the second. Indeed, except for μ close to zero it makes little difference if one neglects the second term entirely. We shall exploit this possibility when analogous but more complicated probabilities enter the discussion below.

Writing the inequality $|S_n| > bn^{1/2}$ in the form $S_n^2/n > b^2$ and observing that S_n^2/n is just the standard log likelihood ratio statistic for testing $\mu = 0$ suggest a generalization of this test to comparison of three treatments. Such a generalization was discussed by Siegmund [(1980), (1985), Chapter V]. That test stops sampling as soon as it becomes clear that there is some treatment effect. It is also possible to give confidence regions for the treatment effect vector and hence a more precise analysis of the nature of the treatment effect.

Nevertheless, the existing theory leads to the awkward possibility that one might terminate the trial early because one treatment has proved inferior to the other two and yet be unable to make a choice between the other two because insufficient data have accumulated to make that discrimination possible. An obvious proposal is to continue with a two treatment repeated significance test to compare the remaining treatments. The overall properties of the combined procedure are the subject of this paper.

2. Definition of the test and quantities to be evaluated. Let $Y_{i,j}$ denote the response of the jth subject to the ith treatment ($i = 1, 2, 3$, $j = 1, 2, \ldots$). Assume that the $Y_{i,j}$ are independent and normally distributed with unknown means μ_i and common known variance which without loss of generality can be taken to equal 1. To be specific, suppose that a large mean response is desirable. Initially observations are taken in triples $Y_j = (Y_{1,j}, Y_{2,j}, Y_{3,j})'$, $j = 1, 2, \ldots$. At any stage j, the experimenter can discard one of the treatments as inferior and continue with the other two, in which case the subsequent observations are taken in pairs.

It will be convenient to develop our theoretical results in a different coordinate system. Let C be the orthogonal matrix

$$C = \begin{pmatrix} 6^{-1/2} & 6^{-1/2} & -2 \cdot 6^{-1/2} \\ 2^{-1/2} & -2^{-1/2} & 0 \\ 3^{-1/2} & 3^{-1/2} & 3^{-1/2} \end{pmatrix}$$

and $X_j = CY_j$, $j = 1, 2, \ldots$. Also let $\theta_1 = 6^{-1/2}(\mu_1 + \mu_2 - 2\mu_3)$ and $\theta_2 = 2^{-1/2}(\mu_1 - \mu_2)$, so the expected value of $X_{i,j}$ is θ_i for $i = 1, 2$ and the hypothesis of no treatment effect, that is, $\mu_1 = \mu_2 = \mu_3$, is equivalent to $\theta_1 = \theta_2 = 0$. Let $S_{i,n} = \sum_{j=1}^{n} X_{i,j}$ and $S_n = (S_{1,n}, S_{2,n})'$. The usual χ^2 statistic based on n observations for testing $\theta_1 = \theta_2 = 0$ is $\|S_n\|^2/n$.

In order to convert back to the original coordinate system, it will be convenient to use the notation $X_j^{i_1 i_2} = 2^{-1/2}(Y_{i_1,j} - Y_{i_2,j})$ and $S_n^{i_1 i_2} = \sum_{j=1}^{n} X_j^{i_1 i_2}$. Observe that $X_j^{12} = X_{2,j}$, $X_j^{13} = (3^{1/2}X_{1,j} + X_{2,j})/2$ and $X_j^{23} = (3^{1/2}X_{1,j} - X_{2,j})/2$ are projections of the vector X_j along the rays rotated $\pi/2$, $\pi/6$ and $-\pi/6$, respectively, from the positive x_1 axis. Also note that $S_n^{12} = \max_{k,l}|S_n^{kl}|$ if and only if ω_n, the angle from the positive x_1 axis to S_n, lies in the interval $[\pi/3, 2\pi/3]$. Similar descriptions in terms of ω_n hold for all events of the form $\{S_n^{i_1 i_2} = \max_{k,l}|S_n^{kl}|\}$ and are used without comment below. A picture involving a hexagon centered at the origin is helpful to display these relations.

The test to be studied below is a concatenation of two repeated significance tests: The first to decide if there is some treatment effect and eliminate the worst treatment, and the second to compare the remaining two treatments. It is determined by six parameters: $b_1 \geq c_1 > 0$, $c_2 \in (0, c_1)$, $b_2 \in [c_2, b_1)$ and $m_0 < m$. The first stage of sampling consists of observing $X_1, X_2, \ldots$ until $\min(T_1, m)$, where

$$T_1 = \inf\{n: n \geq m_0, \|S_n\| > b_1 n^{1/2}\}.$$

(i) If $T_1 < m$ and $S_{T_1}^{i_1 i_3} = \max_{k,l} |S_{T_1}^{kl}|$, the treatment i_3 is discarded and sampling of $X_j^{i_1 i_2}$ continues until $\min(T_2, m)$, where

$$T_2 = \inf\{n\colon n \geq T_1,\ |S_n^{i_1 i_2}| > b_2 n^{1/2}\}.$$

If $T_2 < m$ and $S_{T_2}^{i_1 i_2} > b_2 T_2^{1/2}$ or $T_2 \geq m$ and $S_m^{i_1 i_2} > c_2 m^{1/2}$, treatment i_1 is selected as best, whereas i_2 is selected as best if $T_2 < m$ and $S_{T_2}^{i_1 i_2} < -b_2 T_2^{1/2}$ or $T_2 \geq m$ and $S_m^{i_1 i_2} < -c_2 m^{1/2}$. If $T_2 \geq m$ and $-c_2 m^{1/2} \leq S_m^{i_1 i_2} \leq c_2 m^{1/2}$, then no choice is made between i_1 and i_2.

(ii) If $T_1 \geq m$, $\|S_m\| > c_1 m^{1/2}$ and $S_m^{i_1 i_3} = \max_{k,l} |S_m^{kl}|$, then population i_3 is discarded. If also $S_m^{i_1 i_2} > c_2 m^{1/2}$, respectively $S_m^{i_1 i_2} < -c_2 m^{1/2}$, then treatment i_1, respectively i_2 is selected as best; otherwise no choice is made between i_1 and i_2.

(iii) If $T_1 \geq m$ and $\|S_m\| < c_1 m^{1/2}$, the hypothesis of equal treatment effects, $\theta_1 = \theta_2 = 0$, is not rejected.

ASSUMPTION. We assume throughout that $b_2 \leq b_1 \cdot 3^{1/2}/2$.

REMARKS. (i) The assumption implies that if $T_1 \leq m$, then $\max |S_{T_1}^{kl}| \geq b_2 T_1^{1/2}$. Also, if at time T_1 ω_{T_1} is sufficiently close to $\pi/3$, $-\pi/3$ or π, two treatments are simultaneously eliminated and the test terminates.

(ii) The initial stage of sampling might reasonably be defined in terms of $\max_{k,l} |S_n^{kl}|$ instead of $\|S_n\|$. Since this alternative produces simulated outcomes similar to those reported in the following section and seems more difficult to analyze theoretically, it has not yet been studied in detail.

(iii) There is no special reason that the truncation point m for the second stage should equal the maximum number of triples for the first stage. On the contrary, if $T_1 < m$, one might argue that the $3(m - T_1)$ remaining subjects potentially available for the first stage of sampling should be available for comparing the two treatments remaining at the second stage. In this case the second stage would terminate at $\min[T_2, (3m - T_1)/2]$. In many clinical trials, however, there is a delay between treatment assignment and response, although the simple model described above ignores this possibility. In such cases the intake period will often terminate before the first stage of sampling and no additional subjects will be available for the second stage. See Siegmund [(1985), Section V.6] for a related discussion in the context of two treatment trials involving survival data. We shall return to this and related issues in Section 3.

In the following discussion $\mu_1 \geq \mu_2 \geq \mu_3$, so treatment 1 is at least as good as the other two. Qualitatively different, interesting cases are (a) $\mu_1 = \mu_2 = \mu_3$, (b) $\mu_1 = \mu_2 > \mu_3$ and (c) $\mu_1 > \mu_2 > \mu_3$. In case (a) a Type I-1 error occurs if the first stage of sampling leads to rejection of the hypothesis of no treatment effect that $\theta_1 = \theta_2 = 0$. Its probability is just the significance level of the

repeated significance test defined by b_1, c_1, m_0 and m, that is,

$$P_0\{T_1 < m\} + P_0\{T_1 \geq m, \|S_m\| > c_1 m^{1/2}\}$$
$$= P_0\{\|S_m\| > c_1 m^{1/2}\} + P_0\{T_1 < m, \|S_m\| < c_1 m^{1/2}\},$$

for which Siegmund [(1985), Section V.3] gives an approximation. In cases (b) and (c) a Type II-1 error occurs if the test fails to reject the hypothesis that $\theta_1 = \theta_2 = 0$. The error probability involves the complementary events to those of the preceding case and can again be approximated by reference to the results of Siegmund [(1985), Section V.3]. These approximations involving the first stage of the test are reasonably accurate, as one can see by comparing them to simulations given by Siegmund (1980) and exact numerical calculations in Jennison and Turnbull (1991).

In case (b), where $\mu_1 = \mu_2$ and hence $\theta_2 = 0$, one can also make the Type I-2 error of eliminating treatment 1 or treatment 2. An exact description of this event is complicated. Assume $\theta^0 = (\theta_1, 0)$ with $\theta_1 > 0$. The probability of incorrectly eliminating treatment 1 or 2 is [cf. (1)]

$$\begin{aligned}
&P_{\theta^0}\{T_1 < m, |\omega_{T_1}| < \pi/3, |S_{2,m}| > c_2 m^{1/2}\} \\
&\quad + P_{\theta^0}\{T_1 \geq m, \|S_m\| > c_1 m^{1/2}, |S_{2,m}| > c_2 m^{1/2}\} \\
&\quad + P_{\theta^0}\{T_1 < m, |\omega_{T_1}| < \pi/3, T_2 < m, |S_{2,m}| \leq c_2 m^{1/2}\} \\
&\quad + P_{\theta^0}\{T_1 < m, |\omega_{T_1}| > \pi/3\} \\
&= P_{\theta^0}\{|S_{2,m}| > c_2 m^{1/2}\} \\
&\quad + P_{\theta^0}\{T_1 < m, |\omega_{T_1}| < \pi/3, T_2 < m, |S_{2,m}| \leq c_2 m^{1/2}\} \\
&\quad + P_{\theta^0}\{T_1 < m, |\omega_{T_1}| > \pi/3, |S_{2,m}| < c_2 m^{1/2}\} \\
&\quad - P_{\theta^0}\{T_1 \geq m, \|S_m\| < c_1 m^{1/2}, |S_{2,m}| > c_2 m^{1/2}\}.
\end{aligned} \tag{2}$$

For large values of θ_1 the last term is negligible. If c_2 is small compared to b_2, so the first term on the right-hand side of (2) is the dominant one, the sum of the first three terms on the right-hand side of (2) roughly equals the significance level of the two treatment repeated significance test defined by b_2, c_2, m_0 and m [cf. (1)]. For θ_1 closer to zero, the probability (2) can be somewhat smaller than the significance level of this two treatment test. Section 3 contains a numerical example and Section 4 gives related theoretical calculations.

In case (c), where $\theta_1 > 0$ and $\theta_2 > 0$, one is interested in the Type II-2 error of failure to conclude that treatment 1 is preferred to the others. We shall neglect the possibility that at $T_1 < m$ treatment 1 is incorrectly eliminated, which is an event of very small probability unless $\|\theta\|$ is close to zero. If we also neglect some difficult to compute terms similar to the usually negligible final term on the right-hand side of (1), it is easy to obtain an approximation to the

probability of correctly selecting treatment 1:

$$P_\theta\{\|S_m\| > c_1 m^{1/2}, \min(S_m^{12}, S_m^{13}) > c_2 m^{1/2}\} + P_\theta\{T_1 < m, \|S_m\| < c_1 m^{1/2}, \min(S_m^{12}, S_m^{13}) > c_2 m^{1/2}\}. \tag{3}$$

The final inequality inside the braces in each term of (3) specifies that $\max |S_m^{kl}|$ is attained by either S_m^{12} or by S_m^{13} and if attained by say S_m^{12}, then $S_m^{13} > c_2 m^{1/2}$, so both treatments 2 and 3 are eliminated in favor of treatment 1. Typical events neglected in (3) involve sample paths for which $T_1 < m$, $T_2 < m$, but the statistic for the treatments sampled in the second phase, say S_m^{12} is less than $c_2 m^{1/2}$. The probability of these events and also the second term in (3) are usually very small and can be neglected.

It is easy to approximate each term in (3). The first can be evaluated exactly by double numerical integration. For the second term one can write

$$P_\theta\{T_1 < m, \|S_m\| < c_1 m^{1/2}, \min(S_m^{12}, S_m^{13}) > c_2 m^{1/2}\} = E_\theta\big[P_0\{T_1 < m | S_m\}; \|S_m\| < c_1 m^{1/2}, \min(S_m^{12}, S_m^{13}) > c_2 m^{1/2}\big], \tag{4}$$

apply Theorem 9.54 of Siegmund (1985) to evaluate approximately the conditional probability, and perform a double numerical integration. The same method applies to the final term on the right-hand side of (2), which usually can be neglected.

To assess the expected sample size of the test defined above we consider

$$E_\theta[\min(T_1, m)] \tag{5}$$

and

$$E_\theta[\min(T_2, m)]. \tag{6}$$

The quantity (5) involves only the first stage of sampling. An approximation for it has been given by Siegmund [(1985), Section V.3]. One hopes that at least when θ_2 is substantially greater than zero the expected sample size in (6) is not much greater than for the two treatment test to compare treatments 1 and 2. However, the second stage of sampling is delayed by the requirement that it not begin until the termination of the first stage. Hence (6) always exceeds (5), which may itself in some cases exceed the expected sample size for a trial involving only two treatments. Numerical examples are given in Section 3 and an analytic approximation for (6) in Section 4.

3. Monte Carlo results. This section gives the results of a Monte Carlo experiment to see how the procedure defined in Section 2 performs. Only a single test is considered; but the extensive numerical studies of repeated significance tests for comparing two treatments strongly suggest that the basic conclusions are applicable over a broad range of sample sizes. For our example $m_0 = 10$ and $m = 50$, and observations, that is, triples or pairs, are taken one at a time. One can expect similar results for a group sequential test where the data are examined a maximum of five times, and observations are taken 10 at

a time; or after an appropriate change of scale, where observations are taken k at a time for essentially any k.

The calculations of the preceding section indicate how the parameters b_1, c_1, b_2 and c_2 should be chosen. In order to control the Type I-1 error probability, b_1 and c_1 must be chosen so that the significance level of the three treatment comparison which constitutes the first phase of the test is about equal to the desired value of α. For $\alpha = 0.05$ the approximation given by Siegmund [(1985), Chapter V] (backed up by a Monte Carlo experiment to confirm its accuracy) shows that one possibility is $b_1 = 3.5$ and $c_1 = 2.5$. The Type I-2 error probability depends to some extent on b_1 and c_1, but when $\theta^0 = (\theta_1, 0)'$ with θ_1 large enough that $P_{\theta^0}\{T_1 < m\}$ is close to one, the Type I-2 error probability depends primarily on b_2 and c_2, which must be chosen so that the significance level of a two treatment repeated significance test is approximately α. For $\alpha = 0.05$ it suffices to take $b_2 = 2.92$ and $c_2 = 2.05$. As assumed above, b_2 is smaller than $b_1 \cdot 3^{1/2}/2$. This allows T_2 to equal T_1 in the case that two treatments behave almost identically while being outperformed by the third.

Table 1 contains the results of a Monte Carlo experiment. The following notation denotes the contents of each column: p_1 is the probability of rejecting the global hypothesis of no treatment effect; p_2 denotes the probability of incorrectly eliminating either treatment 1 or treatment 2 in those rows where $\theta_2 = 0$ and denotes the probability of correctly selecting treatment 1 as best in those rows where $\theta_2 > 0$; E_1 is the expected number of triples in the first phase of the test, $E_\theta[\min(T_1, m)]$; E_2 denotes $E_\theta[\min(T_2, m)]$; p_2^* and E_2^* denote the power and expected sample size, respectively, of the repeated significance test for comparing treatments 1 and 2 defined by m_0, m, b_2 and c_2. The three entries immediately under each row in which $\theta_2 = 0$ involve vectors θ oriented at angles of $\pi/3$, $\pi/4$ and $\pi/6$ with respect to the positive θ_1 axis and having the same norm as the value of θ_1 in the preceding row where $\theta_2 = 0$. The values of both p_1 and E_1 are the same for each of these four entries. When $\arg \theta = \pi/3$, treatments 2 and 3 are equally effective. This is the situation most favorable to selecting treatment 1 for a given value of $\|\theta\|$. Of course, for a given value of θ_2 the situation becomes more favorable with increasing θ_1.

Entries not in parentheses are Monte Carlo estimates based on 9999 repetitions in cases where $\theta_2 = 0$ and 2500 repetitions when it does not. Entries in parentheses are computed from theoretical approximations. For p_1, p_2^*, E_1 and E_2^* the approximations given by Siegmund [(1985), Chapters IV and V] have been used. For p_2 the approximation developed in (3) and (4) has been used in those cases, where $\theta = (\theta_1, \theta_2)'$ with θ_2 not equal to 0. The approximation used for E_2 is given in Section 4, where the problem of approximating p_2 in the case $\theta = (\theta_1, 0)'$ is also discussed. Since the present problem is substantially more complex than those considered before, the new approximations seem to be reasonably accurate and to convey a correct impression of the properties of the test under consideration here.

Clearly the test under discussion cannot perform as well as a two treatment repeated significance test with the same m_0, m, $b = b_2$ and $c = c_2$. An important question to be considered when examining Table 1 is whether it

TABLE 1

Operating characteristics $m_0 = 10$, $m = 50$, $b_1 = 3.5$, $c_1 = 2.5$, $b_2 = 2.92$, $c_2 = 2.05$; *Type I-1 error probability is* 0.05[1]

θ_1	θ_2	p_1	p_2	p_2^*	E_1	E_2	E_2^*
0.70	0.00	(0.995)	0.050 (0.049)	(0.050)	23.3 (22.7)	49.2	
0.35	0.61	(0.995)	0.982 (0.977)	(0.989)	23.3 (22.7)	28.7 (26.6)	(21.9)
0.50	0.50	(0.995)	0.942 (0.930)	(0.936)	22.9 (22.7)	32.2 (32.3)	(29.7)
0.61	0.35	(0.995)	0.670 (0.664)	(0.682)	22.8 (22.7)	41.2	(38.2)
0.60	0.00	(0.972)	0.048 (0.048)	(0.050)	29.4 (29.5)	49.4	
0.30	0.52	(0.972)	0.906 (0.908)	(0.952)	28.9 (29.5)	35.2 (33.2)	(28.1)
0.42	0.42	(0.972)	0.822 (0.807)	(0.831)	29.4 (29.5)	38.2 (39.4)	(36.0)
0.52	0.30	(0.972)	0.552 (0.527)	(0.564)	29.5 (29.5)	44.0	43.1
0.50	0.00	(0.890)	0.047 (0.048)	(0.050)	35.8 (36.8)	49.6	
0.25	0.43	(0.890)	0.756 (0.741)	(0.848)	36.3 (36.8)	41.8 (39.3)	(35.3)
0.35	0.35	(0.890)	0.629 (0.636)	(0.682)	35.9 (36.8)	43.1 (44.4)	(38.2)
0.43	0.25	(0.890)	0.400 (0.383)	(0.414)	35.8 (36.8)	46.2	
0.40	0.00	(0.701)	0.046 (0.046)	(0.050)	41.6	49.6	
0.20	0.35	(0.701)	0.503 (0.057)	(0.682)	42.0	45.8	(38.2)
0.28	0.28	(0.701)	0.398 (0.414)	(0.495)	42.3	46.9	43.9
0.35	0.20	(0.701)	0.260 (0.247)	(0.282)	41.5	47.9	
0.30	0.00	(0.451)	0.041 (0.041)	(0.050)	46.3	49.6	
0.15	0.26	(0.451)	0.264 (0.252)	(0.441)	46.3	48.5	45.1
0.21	0.21	(0.451)	0.214 (0.209)	(0.313)	46.2	48.6	
0.26	0.15	(0.451)	0.141 (0.132)	(0.178)	46.1	49.2	

[1] p_1 is the probability of rejecting the global hypothesis of no treatment effect; p_2 denotes the probability of incorrectly eliminating Treatment 1 or 2 in those rows where $\theta_2 = 0$ and denotes the probability of correctly selecting Treatment 1 in those rows where $\theta_2 > 0$; $E_1 = E_\theta[\min(T_1, m)]$; $E_2 = E_\theta[\min(T_2, m)]$; p_2^* and E_2^* denote the power and expected sample size of the repeated significance test for comparing Treatments 1 and 2 defined by m_0, m, b_2, c_2.

performs substantially worse. The answer is not completely obvious. In many rows p_2 and E_2 are about equal to p_2^* and E_2^*. But in several rows our test suffers either a substantial loss of power, an increase in the expected sample size, or sometimes both. Large percentage increases in expected sample size tend to occur in the upper part of the table, where $\|\theta\|$ is large, expected sample sizes are small, and the expected number of triples in the first phase of our test may be as large as the expected number of pairs required to compare the two better treatments. A significant loss of power compared to the two treatment test tends to occur lower in the table, where $\|\theta\|$ is small and hence our test may not even enter into its second phase. To some extent these problems are inevitable, although with fine tuning it might be possible to devise a scheme which does a bit better than the one considered here. It seems fair to conclude that our test is not obviously defective, although some improvement may be possible.

Although the preceding comparison of our test with a two treatment repeated significance test seems to be useful for purposes of calibration, it is presumably not the comparison one would make when trying to decide whether

a trial should involve three treatments or only two. Then a more reasonable comparison might be between a three treatment trial employing a maximum of m patients on each of the three treatments for a total of $3m$ patients, who in a two treatment trial could be allocated to only two treatments, so there would be a maximum of $3m/2$ subjects receiving each treatment. The two treatment trial would have much greater power, so one would prefer the three treatment trial only if the third treatment is a credible option. On the other hand, if the choice is between a three treatment trial or consecutive two treatment comparisons, the three treatment trial seems to offer definite advantages, although this issue might be worth more systematic investigation.

As indicated in the preceding section, it is also interesting to suppose that a total of $3m$ subjects are available for the trial and those who do not participate in the first phase are available for the second. More precisely suppose that if the first phase terminates at $T_1 < m$, then $3(m - T_1)$ additional subjects are available for assignment to the remaining treatments of the second phase, which therefore terminates at $\min[T_2, (3m - T_1)/2]$. This will increase the Type I-2 error probability, but only slightly. The probability of correctly selecting the best treatment will also increase, most notably in cases where $\|\theta\|$ is large, so the first phase is quickly terminated, but θ_2 is relatively small, so the extra subjects made available by that early termination are useful in the second phase. If $\|\theta\|$ and θ_2 are both large, the results will be much as in Table 1, because it is usually unnecessary to take more observations before reaching a decision; and if $\|\theta\|$ is small, the results will again be much as in Table 1 because the first phase of the test will usually exhaust the entire group of subjects and very few will remain for the second phase. A numerical example emphasizing parameter values for which one can expect large differences with Table 1 is given in Table 2.

The notation in Table 2 is the same as in Table 1. Since the approximations suggested in Section 4 for the Type I-2 error probabilities appeared to be quite accurate in Table 1, appropriately modified analytic approximations are used in Table 2. The starred columns also give analytic approximations for the two treatment repeated significance test with $m_0 = 10$, $m = 70$, $b = 2.92$ and $c = 2.05$. This value of m would be appropriate in the ideal case that the first

TABLE 2

Operating characteristics when the second phase terminates at $\min[T_2, (3m - T_1)/2]$

θ_1	θ_2	p_2	p_2^*	E_2	E_2^*
0.70	0.00	(0.052)	(0.054)	62.2	
0.61	0.35	0.763	(0.822)	47.2	(50.2)
0.60	0.00	(0.051)	(0.054)	59.4	
0.42	0.42	0.856	(0.933)	40.2	(41.3)
0.52	0.30	0.610	(0.695)	49.1	(52.9)
0.50	0.00	(0.049)	(0.054)	56.3	
0.43	0.25	0.444	(0.540)	50.4	60.6

phase of the three treatment test always terminates at the minimum sample size, $T_1 = 10$, so the maximum sample size at the end of the second phase is $(3 \times 50 - 10)/2 = 70$. The other entries in Table 2 are Monte Carlo estimates based on 2500 repetitions. In comparison with Table 1, our test takes more observations and achieves a nontrivial increase in the probability of correctly selecting the best treatment, although it still falls short of the ideal case represented by the two treatment trial described in the starred columns.

It is interesting to note that the Type I-2 error is not substantially larger than the nominal level of 0.05. In fact, one of the most attractive features of the repeated significance test with respect to the O'Brien–Fleming test is this robustness of the significance level with respect to changes in the maximum sample size m, which is only rarely known before an experiment begins even in two treatment trials. An increase of the maximum sample size from 50 to 70 for an O'Brien–Fleming test would more than double the significance level if no other adjustment is made.

4. Analytic approximations. This section is concerned with analytic approximations for the expected sample size (6) and the Type I-2 error probability given (approximately) in (2).

The approximation given here for (6) is similar in spirit to that given for (5) in Siegmund [(1985), Chapter V]; but it is substantially more complicated and hence is given to less precision. Nevertheless, the results in Table 1 show that it can be quite good.

Our point of departure is the following identity, which may be readily verified:

$$E_\theta[\min(T_2, m)] = E_\theta[\min(T_1, m)] + E_\theta[T_2 - T_1; T_1 \le m] - E_\theta[T_2 - m; T_1 \le m < T_2]. \tag{7}$$

Under the assumption

$$\|\theta\| = \gamma_1 + \Delta_1/m^{1/2}, \quad \text{where } \gamma_1 = b_1/m^{1/2} \tag{8}$$

and Δ_1 are constants as $m \to \infty$ with $\gamma_1 > 0$, an asymptotic expansion for $E_\theta[\min(T_1, m)]$ up to terms which converge to zero as $m \to \infty$ has been given by Siegmund [(1985), page 113]. The approximations for $E_\theta\{\min(T_2, m)\}$ given in Table 1 are the sum of that approximation and the approximations given below for the second and third terms on the right-hand side of (7), which under appropriate assumptions are both of order $m^{1/2}$.

For $\mu_1 \gg \mu_2 \ge \mu_3$, up to a negligibly small error the second term on the right-hand side of (7) equals

$$E_\theta\Big(T_2 - T_1; T_1 < m, S_{T_1}^{13} = \max_{i,j} |S_{T_1}^{ij}|\Big) + E_\theta\Big(T_2 - T_1; T_1 < m, S_{T_1}^{12} = \max_{i,j} |S_{T_1}^{i,j}|\Big). \tag{9}$$

When μ_2 is substantially larger than μ_3, the second term in (9) is negligible,

but when μ_2 and μ_3 are approximately equal, so are the two terms in (9). If $\mu_2 = \mu_3$, that is, $\theta_2 = \|\theta\| \cdot 3^{1/2}/2$, the two terms are exactly equal.

To analyze (9) we assume that

$$\theta_2 = \gamma_2 + \Delta_2/m^{1/2}, \tag{10}$$

where $\gamma_2 = b_2/m^{1/2}$, and also that

$$\theta_2 = \|\theta\| \cdot 3^{1/2}/2 + O(m^{-1/2}). \tag{11}$$

Note that conditions (8) and (10) imply that

$$\theta_2 = b_2\|\theta\|/b_1 + \Delta_3/m^{1/2}, \tag{12}$$

where $\Delta_3 = \Delta_2 - \gamma_2\Delta_1/\gamma_1$.

REMARK. Condition (12) implies that $T_2 - T_1$ is of order $m^{1/2}$ in probability. If (12) holds, conditions (8) and (10) could be replaced by the assumptions that $\|\theta\| > \gamma_1$ and $\theta_2 > \gamma_2$ in the calculations given below. Then the third term on the right-hand side of (7) would be asymptotically negligible. For the entries in Table 1 these conditions are often more appropriate than those given above. The condition (11) is also unnecessarily restrictive. It implies that the two terms in each of the displays (9) and (20) are of the same order of magnitude. If (11) does not hold, the second terms in (9) and (20) can be neglected.

Since the following calculations involve a large number of technical details which for similar problems have been treated elsewhere [e.g., Siegmund (1985)], many details have been omitted here.

From the assumptions that $b_2 \le b_1 \cdot 3^{1/2}/2$, it follows that if $T_1 = n$ and $S_{2,n} < b_2 n^{1/2}$, then except for an event of exponentially small probability $S_n^{13} > S_n^{12}$, and hence except for an exponentially small error the first term in (9) equals

$$\sum_{m_0 \le n \le m} \int_{x>0} P_\theta\{T_1 = n,\ S_{2,n} \in b_2 n^{1/2} - dx\} E_\theta \tau_n(x), \tag{13}$$

where $\tau_n(x) = \inf\{k\colon S_{2,k} \ge b_2(n+k)^{1/2} - b_2 n^{1/2} + x\}$. Assume $n \sim (b_1/\|\theta\|)^2$ and $m^{1/8} < x < m^{5/8}$. Since $T_1 \sim (b_1/\|\theta\|)^2$ in probability, as in Proposition 4.27 of Siegmund (1985) one may show that these ranges of n and x make the dominant contribution to (13) when (8) and (10) hold, and in these ranges

$$E_\theta[\tau_n(x)] \sim x/(\theta_2 - b_2/2n^{1/2}) \sim x/(\theta_2 - b_2\|\theta\|/2b_1). \tag{14}$$

A likelihood ratio calculation [Siegmund (1985), page 114] shows that for any stopping rule defined in terms of $\|S_i\|$, $i = 1, 2, \ldots$, in particular for T_1,

$$P_\theta\{S_{2,T_1} \in d\xi \,|\, T_1, \|S_{T_1}\|\} = \frac{\exp\{\|\theta\|\,\|S_{T_1}\|\cos[\sin^{-1}(\xi/\|S_{T_1}\|) - \sin^{-1}(\theta_2/\|\theta\|)]\}\, d\xi}{2\pi I_0(\|\theta\|\,\|S_{T_1}\|)(\|S_{T_1}\|^2 - \xi^2)^{1/2}}, \tag{15}$$

where I_0 denotes the usual modified Bessel function.

Since $T_1/b_1^2 \sim \|\theta\|^{-2}$ in probability and $\|S_{T_1}\| = b_1 T_1^{1/2} + R$, where R is stochastically bounded (and in fact converges in distribution), by (12), the asymptotic relation $I_0(v) \sim \exp(v)/(2\pi v)^{1/2}$ as $v \to \infty$, and some extensive Taylor series expansions, one finds that for $m^{1/8} < x < m^{5/8}$, on an event having probability converging to 1,

$$
\begin{aligned}
&P_\theta\{S_{2,T_1} \in b_2 T_1^{1/2} - dx | T_1, \|S_{T_1}\|\} \\
&\qquad \sim \phi\left[\gamma_1\Delta_3/\theta_1 + x\|\theta\|^2/\left(m^{1/2}\gamma_1\theta_1\right)\right] dx\|\theta\|^2/\left(m^{1/2}\gamma_1\theta_1\right),
\end{aligned} \tag{16}
$$

where $\Delta_3 = \Delta_2 - \gamma_2\Delta_1/\gamma_1$ and ϕ denotes the standard normal probability density function.

Finally, it is well known that T_1 and $R = \|S_{T_1}\| - b_1 T_1^{1/2}$ are asymptotically independent [cf. Siegmund (1985), Theorem 9.17], and that when (8) holds,

$$
P_\theta\{T_1 \le m, \|S_m\| \le b_1 m^{1/2}\} = O(m^{-1/2}), \tag{17}
$$

so

$$
\begin{aligned}
P_\theta\{T_1 \le m\} &= P_\theta\{\|S_m\| > b_1 m^{1/2}\} + O(m^{-1/2}) \\
&= 1 - \Phi\left(b_1 - \|\theta\| m^{1/2}\right) + O(m^{-1/2}) \\
&= 1 - \Phi[-\Delta_1] + O(m^{-1/2}),
\end{aligned} \tag{18}
$$

where Φ denotes the standard normal distribution function.

Substitution of (14), (16) and (18) into (13) and integration yield the following result.

PROPOSITION 1. *Assume* (8), (10) *and* (11) *hold. Let* $\Delta_3 = \Delta_2 - \gamma_2\Delta_1/\gamma_1$, $\eta = \Delta_3\gamma_1/\theta_1$ *and* $\sigma = \gamma_1\theta_1/\|\theta\|^2$. *Then as* $m \to \infty$,

$$
\begin{aligned}
&E_\theta\left(T_2 - T_1; T_1 < m, S_{T_1}^{13} = \max|S_{T_1}^{ij}|\right) \\
&\qquad \sim P_\theta\{T_1 \le m\}\sigma m^{1/2}[\phi(\eta) - \eta\Phi(-\eta)]/(\theta_2 - b_2\|\theta\|/2b_1),
\end{aligned} \tag{19}
$$

where $P_\theta\{T_1 \le m\}$ *is given approximately by* (18).

Now consider the final expectation in (7). Under the assumption that $\mu_1 \gg \mu_2 \ge \mu_3$, up to an exponentially small error this term equals

$$
\begin{aligned}
&E_\theta\left\{T_2 - m; T_1 \le m < T_2, S_{T_1}^{13} = \max|S_{T_1}^{ij}|\right\} \\
&\qquad + E_\theta\left\{T_2 - m; T_1 \le m < T_2, S_{T_1}^{12} = \max|S_{T_1}^{ij}|\right\}.
\end{aligned} \tag{20}
$$

As above, if μ_2 is substantially larger than μ_3 the second term in (20) is negligible compared to the first and can be ignored. If μ_2 and μ_3 are about equal, both terms can be approximated by the following argument. Of course, if $\mu_2 = \mu_3$, the two terms are exactly equal. By (17) the first term in (20) has

the same asymptotic behavior as

$$\int_{x>0} P_\theta\Big\{\|S_m\| > b_1 m^{1/2},\ T_2 > m,\ S_{T_1}^{13} = \max|S_{T_1}^{ij}|,\ S_{2,m} \in b_2 m^{1/2} - dx\Big\} E_\theta[\tau_m(x)]. \tag{21}$$

For x in the critical range $m^{1/8} < x < m^{5/8}$, we can ignore the constraint $S_{T_1}^{13} = \max|S_{T_1}^{ij}|$, since the conditional probability

$$P_\theta\Big\{T_1 < m,\ S_{T_1}^{12} = \max|S_{T_1}^{ij}| \,|S_{2,m} = b_2 m^{1/2} - x\Big\}$$

$$\le P_\theta\{S_{2,n} > b_2 n^{1/2} \text{ for some } m_0 \le n < m | S_{2,m} = b_2 m^{1/2} - x\} \to 0$$

as $m \to \infty$ [(cf. Siegmund (1985), page 200]. Similarly we have $P_\theta\{T_2 < m, |S_{2,m}| < b_2 m^{1/2}\} = O(m^{-1/2})$, so we can also ignore the constraint $T_2 > m$ in (21). After a lengthy calculation one sees that

$$\begin{aligned} &P_\theta\{\|S_m\| > b_1 m^{1/2},\ S_{2,m} \in b_2 m^{1/2} - dx\} \\ &\sim \phi(\Delta_2 + x/m^{1/2}) \\ &\quad \times\{1 - \Phi[(-\Delta_1\|\theta\| + \Delta_2\theta_2 + \theta_2 x m^{-1/2})/\theta_1]\}\, dx/m^{1/2}. \end{aligned} \tag{22}$$

Finally, substitution of (14) and (22) into (21) yields an expression which is easily computed numerically. The final approximation is summarized in the following proposition.

PROPOSITION 2. *Assume conditions* (8), (10) *and* (11). *Then as* $m \to \infty$,

$$\begin{aligned} &E_\theta\Big(T_2 - m; T_1 \le m < T_2,\ S_{T_1}^{13} = \max|S_{T_1}^{ij}|\Big) \\ &\sim m^{1/2}(\theta_2 - \gamma_2/2)^{-1} \\ &\quad \times \int_0^\infty y\phi(\Delta_2 + y)\{1 - \Phi[(-\Delta_1\|\theta\| + \Delta_2\theta_2 + \theta_2 y)/\theta_1]\}\, dy. \end{aligned} \tag{23}$$

REMARK. By scrutinizing the argument given above, one can see that there is in principle no impediment to calculating a term of order one in an asymptotic expansion of $E_\theta\{\min(T_2, m)\}$. Unfortunately the result would be extremely complicated and involve some terms which seem to give insight into the situation and others which just appear in the calculation. Since there are other approaches to an approximation for $E_\theta\{\min(T_2, m)\}$, which might produce useful results over a wider range of parameters than the approach taken here, it seems reasonable to regard the above results as demonstrating the possibility of a theoretical analysis, but by no means providing the final word. On the other hand, in spite of the crudeness of the preceding analysis it gives interesting insight into the structure of the expected sample size.

Now consider the Type I-2 error probability given (approximately) by (2). The loose heuristic argument in Section 2 indicates that for large θ_1 that probability should be only slightly less than the significance level of the two treatment repeated significance test defined by b_2, c_2, m_0 and m. Since this reasoning is substantiated by the Monte Carlo results reported in Table 1, it seems interesting to try to provide a more precise explanation; and since the next to last term in (2) seems difficult to evaluate, we shall replace the original probability by the analogous one for Brownian motion. Although the approximation for the new problem cannot be expected to provide numerically accurate results for the original, an ad hoc continuity corrected version does, as we shall see below.

The first term on the right side of (2) is the principal one when c_2 is small compared to b_2. It also makes the principal contribution to the significance level of a two treatment test, and it does not change if the process $\{S_n\}$ is replaced by Brownian motion. As indicated in Section 2, the fourth term in (2) can be approximated by the same method used in (4), but is usually negligible unless θ_1 is small.

To study the behavior of the second term, let $W_t = (W_{1,t}, W_{2,t})'$ be two dimensional Brownian motion with drift $\theta = (\theta_1, \theta_2)'$ per unit time. Let

$$\tau_1 = \inf\{t: t \geq m_0, \|W_t\| > b_1 t^{1/2}\}$$

and

$$\tau_2 = \inf\{t: t \geq \tau_1, |W_{2,t}| > b_2 t^{1/2}\}.$$

Let $\omega_t = \arg(W_t)$. Assume $\theta^0 = (\theta_1, 0)'$ with $\theta_1 > 0$ and consider

$$P_{\theta^0}\{\tau_1 \leq \tau_2 < m, |\omega_{\tau_1}| < \pi/3, |W_{2,m}| \leq c_2 m^{1/2}\}$$

$$= \int_{[m_0, m)} \int_{|\omega| < \pi/3} P_{\theta^0}\{\tau_1 \in dt, \omega_{\tau_1} \in d\omega\} \tag{24}$$

$$\times P_{\theta^0}\{\tau_2 < m, |W_{2,m}| \leq c_2 m^{1/2} | \tau_1 = t, \omega_{\tau_1} = \omega\}.$$

Assume that $m \to \infty$ with $b_2/m^{1/2} = \gamma_2 > 0$, $m_0 = t_0 m$, $c_2/m^{1/2}$ a constant in $(0, \gamma_2]$, and $t/m \in (t_0, 1)$ bounded away from the endpoints. It is shown in the Appendix that

$$P_{\theta^0}\{\tau_2 < m, |W_{2,m}| \leq c_2 m^{1/2} | \tau_1 = t, W_{2,\tau} = \xi\}$$

$$\sim (b_2 t^{1/2}/\xi) \exp\left[-(b_2^2 - \xi^2/t)/2\right], \tag{25}$$

provided there exists $\varepsilon > 0$ such that $(1 + \varepsilon) b_2^2/c_2 m^{1/2} < \xi/t < b_2/t^{1/2}$, and is of smaller order of magnitude if $\xi/t < (1 - \varepsilon) b_2^2/c_2 m^{1/2}$. Known results cited

in the Appendix strongly suggest that

$$P_{\theta^0}\{\tau_1 \in dt,\ \omega_{\tau_1} \in d\omega\}$$

$$\sim \tfrac{1}{2}b_1^2 \exp\left[-\tfrac{1}{2}b_1^2 + \theta_1 b_1 t^{1/2}\cos\omega - \tfrac{1}{2}\theta_1^2 t\right] d\omega\, dt/(2\pi t). \tag{26}$$

Substitution of (25) with $\xi = b_1 t^{1/2}\sin\omega$ and (26) into (24) yields the approximation

$$P_{\theta^0}\{m_0 < \tau_1 < \tau_2 < m,\ |\omega_{\tau_1}| < \sin^{-1}(b_2/b_1),\ |W_{2,m}| \le c_2 m^{1/2}\}$$

$$\sim 2b_2\phi(b_2)b_1\int\int x^{-1}\phi(x - b_1\cos\omega)\, dx\, d\omega/\sin\omega, \tag{27}$$

where the range of integration is $\delta m_0^{1/2} < x < \delta b_1 c_2 m^{1/2}\sin\omega/b_2^2$, $b_2^2 m_0^{1/2}/(b_1 c_2 m^{1/2}) < \sin\omega < b_2/b_1$. The approximation (27) omits several boundary cases in (24) which can make important contributions for some configurations of the parameters. Similar arguments yield

$$P_{\theta^0}\{\tau_1 = m_0,\ |\omega_{m_0}| < \pi/3,\ m_0 < \tau_2 < m,\ |W_{2,m}| \le c_2 m^{1/2}\}$$

$$\sim 2b_2\phi(b_2)\int\int\phi(r\cos\omega - \theta_1 m_0^{1/2})(\sin\omega)^{-1}\, dr\, d\omega, \tag{28}$$

where the range of integration is $b_2^2 m_0^{1/2}/(c_2 r m^{1/2}) < \sin\omega < b_2/r$, $b_1 < r$; and

$$P_{\theta^0}\{m_0 < \tau_1 = \tau_2 < m,\ |\omega_{\tau_1}| < \pi/3,\ |W_{2,m}| \le c_2 m^{1/2}\}$$

$$\sim (2\pi)^{-1/2} b_1^2 \phi(b_1)$$

$$\times \int_{m_0}^{m}\int_{\sin^{-1}(b_2/b_1)<\omega<\pi/3} t^{-1}\exp\left(\theta_1 b_1 t^{1/2}\cos\omega - \theta_1^2 t/2\right)$$

$$\times \Phi\left(\frac{c_2 m^{1/2} - b_1 t^{1/2}\sin\omega}{(m-t)^{1/2}}\right) d\omega\, dt. \tag{29}$$

The probability (29) is closely related to the third term on the right-hand side of (2), and the right-hand side of (29) integrated over values of $\omega > \pi/3$ yields an approximation to that term. Finally, it is easy to evaluate $P_{\theta^0}\{\tau_1 = \tau_2 = m_0,\ |W_{2,m}| \le c_2 m^{1/2}\}$ exactly as a double integral, but since this probability is usually negligible compared to (28) and (29) it is omitted here. The sum of (27), (28) and (29) gives an approximation to (24).

The preceding approximations cannot be expected to be quantitatively accurate for problems in discrete time, but there is a simple modification which appears to be. For the two treatment repeated significance test with

stopping rule $\min(\tau, m)$, where

$$\tau = \inf\{t: t \geq m_0, |W_{2,t}| \geq b_2 t^{1/2}\},$$

one has the approximation [Siegmund (1985), Chapter IV]

$$(30) \qquad P_{\theta^0}\{\tau < m, |W_{2,m}| \leq c_2 m^{1/2}\} \sim 2b_2\phi(b_2)\log[mc_2^2/m_0 b_2^2]^{1/2},$$

which for $m_0 = 10$, $m = 50$, $b_2 = 2.92$ and $c_2 = 2.05$ equals 0.015. The analogous probability in the discrete case is about 0.01 or about $2/3$ as large. The discrete time approximation [Siegmund (1985), page 200] can be regarded as the approximation in (30) multiplied by a factor correcting for excess over the boundary. Since this correction factor changes rather slowly as a function of the various parameters in the problem, one suspects that the same correction factor might be roughly correct for a class of related problems, in particular those discussed above.

The sum of (27)–(29) with the range of integration in (29) extended to include values of $\omega > \pi/3$ gives for $\theta_1 = 0.7$, 0.6, 0.5, 0.4 and 0.3 the values 0.013, 0.012, 0.011, 0.0094 and 0.0074, respectively. If these figures are multiplied by $2/3$ as a correction for discrete time and then added to the principal term on the right-hand side of (2), $2[1 - \Phi(2.05)] \cong 0.040$, they predict quite well the Monte Carlo estimates in Table 1, except for $\theta_1 = 0.3$. Actually for θ_1 as small as 0.3 or 0.4 the final term on the right-hand side of (2) is not negligible, but approximately equals 0.004 or 0.001, respectively. Including this term makes our overall approximation quite good at $\theta_1 = 0.3$ as well. The final continuity corrected approximations are given in Table 1. Simulations for various parameter values, in particular for group sequential tests where sample sizes are nominally very small and corrections for discrete time important, show that this ad hoc correction works well over a broad range of parameter values. A more satisfactory, although still not completely justifiable approximation has recently been obtained by Betensky (1992) and will appear elsewhere.

It is also easy to modify the approximations (27)–(29) so that they apply to the procedure studied in Table 2, where the test is truncated at $T_3 = [3m - \min(T_1, m)]/2$. However, in that case the dominant contribution to the Type I-2 error probability corresponding to the first term on the right-hand side of (2) is

$$P_{\theta^0}\{|W_{2,T_3}| > c_2 T_3^{1/2}\}.$$

This probability does not exactly equal $2[1 - \Phi(c_2)]$, although some simple manipulations suggest it should be very close. The values given in Table 2 for the overall error probability are the sum of $2[1 - \Phi(c_2)]$ and the sum of the modified (27)–(30) corrected as above for discrete time. A Monte Carlo experiment not reported here showed these approximations to be quite accurate.

It seems an interesting, but difficult problem to give asymptotic approximations to the integrals appearing in (27)–(29), which would give sufficient accuracy that one would prefer to use them in lieu of numerical calculations.

5. Discussion and open problems. There remain a number of problems which should be addressed before one feels comfortable with the procedures of this paper. The most obvious is to develop analogous methods for censored survival data. There is by now considerable literature in the two treatment case showing that the simple model discussed here, which involves normally distributed immediately available responses of known variability, provides a good approximate model for censored survival data analyzed, for example, by the proportional hazards model. However, an important aspect of the related large sample theory is a data dependent change of the time scale, for example, Sellke and Siegmund (1983), which adapts automatically to different than anticipated accrual rate, baseline hazard rate, and amount of censoring. In the case of three treatments there are at least two time scales and one cannot devise a single time change which works for all of them simultaneously, except by making somewhat stronger assumptions than the two treatment case requires. The issue is less significant when the treatment effects are all roughly equal, but becomes more important if the treatment effects are radically different and hence information about them accumulates at different rates. See Betensky (1992).

The procedures discussed in this paper are symmetric in the labeling of the treatments. There are various reasons that some treatments, notably a standard treatment or placebo, may be regarded differently from the others. For example, there is usually a presumption that a new treatment is more expensive or more toxic, perhaps both, than a standard and should be adopted only if it shows definite advantages in treatment. Thus the standard treatment need not prove itself superior, but only not inferior to new treatments. For two treatment trials there are easy modifications to allow for early termination of the test if it appears early on that the data are consistent with the hypothesis of no treatment effect. For example, Siegmund (1986) shows that the theory developed for a symmetric repeated significance test can be applied to a natural asymmetric modification of the test which is designed for early termination if the null hypothesis appears to be true. In the case of three treatments it is not obvious how to modify the stopping rule of the first phase of the test we have studied in order to terminate the test quickly in the case neither experimental treatment seems definitely superior to the standard. On the other hand, one of the principal reasons given for wanting to terminate a test if the experimental treatment does not display a definite advantage is to begin a new trial with a new experimental treatment. Insofar as the methods of this paper may allow one to consider several treatments simultaneously, there may not be as much pressure to get on to the next trial. See Betensky (1992).

There is always the possibility that a treatment is discarded early in the test, and by the end appears to be a serious competitor to the other two, if they fail to live up to their early promise. If the criterion for eliminating a

treatment from consideration is reasonably stringent relative to the maximum sample size, as is the case for the example discussed in detail in this paper, it seems unlikely that there will be enough observations remaining after a treatment is discarded to lead to a situation where it ought to be reconsidered. In fact, it seems reasonable to require the probability of such a reversal to be small, and to calculate this probability as a part of one's evaluation of a given test.

It would also be interesting to see if the methods of this paper can be reasonably extended to trials involving four or more treatments and to attempt to give confidence regions for the treatment effect vector at the end of the test. For this latter problem, the methods of Woodroofe [(1986), (1989)] might be useful.

APPENDIX

The approximations (27), (28) and (29) are direct consequences of (25) and (26). This Appendix contains arguments in support of (25) and (26). Since both are closely related to other well known results, most details are omitted.

Display (26) contains an approximation to the probability density function of T_1. In one dimension a very sharp result analogous to (26) has been given by Jennen (1985), and the method used there probably can be adapted to prove (26). An alternative method for one dimension which probably can be used to prove (26) has been developed by Durbin (1985). A discrete time version of (26) has been given by Woodroofe (1978). Analogous results for the distribution function have been obtained, for example, by Siegmund [(1985), Problem 5.1] (where, however, some exponents depending on the dimension of the process are incorrect). In view of this quite substantial related literature there can be little doubt that (26) holds, so we shall not discuss it further.

The relation (25) follows from a modification of known arguments. Since the basic idea is easily understood, it is explained below; but since the details are similar to those given in several other places, they have been omitted. See, for example, Siegmund [(1985), (1988)].

Let w_t be one dimensional Brownian motion with drift μ and for $b > 0$ let

$$\tau = \inf\{s\colon s > t,\ |w_s| > bs^{1/2}\}.$$

Our point of departure is the following representation, which is a synthesis of Lemma A1 of Siegmund (1988) and Problem 4.13 of Siegmund (1985). The proof proceeds along the lines of Wald's likelihood ratio identity and is omitted. Let $Q = \int P_\mu\, d\mu/(2\pi)^{1/2}$, and let L_t denote the likelihood ratio of w_s, $s \le t$ under Q relative to P_0. Then for any $0 < t < m$, $0 < |\xi| < bt^{1/2}$ and $0 < c < b$,

$$\begin{aligned} P_0\{\tau < m,\ |w_m| < cm^{1/2} | w_t = \xi\} \\ = \int E_\mu\left[1_{(\tau < m)} L_\tau^{-1} P_0\{|w_m| < cm^{1/2} | w_s,\ s \le \tau\} | w_t = \xi\right] \\ \times \exp(\mu\xi - \mu^2 t/2)\, d\mu/(2\pi)^{1/2}. \end{aligned}$$

Since $L_t = t^{-1/2}\exp[w_t^2/(2t)]$ and $|w_\tau| = b\tau^{1/2}$ on $\{t < \tau < m\}$, straightforward substitution yields the following identity.

PROPOSITION 3. *Let* $0 < t < m$, $0 < |\xi| < bt^{1/2}$, $0 < c < b$. *Then*

$$
\begin{aligned}
&P_0\{\tau < m, |w_m| < cm^{1/2} | w_t = \xi\} \\
&\quad = t^{-1/2}\exp\left[-\tfrac{1}{2}(b^2 - \xi^2/t)\right] \\
&\quad\quad \times \int_{-\infty}^{\infty} E_\mu\left[1_{(\tau < m)}\tau^{1/2}P_0(|w_m| < cm^{1/2} | \tau, w_\tau) | w_t = \xi\right] \\
&\quad\quad\quad \times \phi\left[(\mu - \xi/t)t^{1/2}\right]t^{1/2}\,d\mu.
\end{aligned} \tag{31}
$$

Note that the inequality

$$P_0\{\tau < m, |w_m| < cm^{1/2} | w_t = \xi\} \le (m/t)^{1/2}\exp\left[-\tfrac{1}{2}(b^2 - \xi^2/t)\right]$$

follows at once from (31).

Assume now the asymptotic scaling of (25): $b_0 = b/m^{1/2}$, $c_0 = c/m^{1/2}$, $\xi_0 = \xi/m$ and $t_0 = t/m$ are constants different from 0. The integrating measure $\phi[(\mu - \xi/t)t^{1/2}]t^{1/2}\,d\mu$ in (31) behaves asymptotically like a delta function at $\xi/t = \xi_0/t_0$, and hence the integral

$$\sim E_{\xi_0 t_0^{-1}}\left[1_{(\tau < m)}\tau^{1/2}P_0(|w_m| < cm^{1/2} | \tau, w_\tau) | w_t = \xi\right].$$

Under $P_{\xi_0 t_0^{-1}}$, $\tau \sim (b/\xi_0 t_0^{-1})^2$ with probability 1; and the conditional probability asymptotically equals

$$\Phi\left(\frac{cm^{1/2} - b\tau^{1/2}}{(m - \tau)^{1/2}}\right),$$

which converges to 1 if $|\xi_0|t_0^{-1} > b_0^2 c_0^{-1}$ and to 0 if $|\xi_0|t_0^{-1} < b_0^2 c_0^{-1}$. Hence by (31),

$$P_0(\tau < m, |w_m| < cm^{1/2} | w_t = \xi) \sim (bt^{1/2}/\xi)\exp\left[-\tfrac{1}{2}(b^2 - \xi^2/t)\right]$$

or is a smaller order of magnitude according as $|\xi_0| > b_0^2 t_0/c_0$ or $|\xi_0| < b_0^2 t_0/c_0$. This is essentially (25) except for the change of notation.

Acknowledgments. I would like to thank the referee and Associate Editor for their helpful comments and suggestions.

REFERENCES

BETENSKY, R. (1992). A study of sequential procedures for comparing three treatments. Ph.D. dissertation, Stanford Univ..

DURBIN, J. (1985). The first passage density of a Gaussian process to a general boundary. *J. Appl. Probab.* **22** 99–122.

JENNEN, C. (1985). Second-order approximations for Brownian first exit distributions. *Ann. Probab.* **13** 126–144.

JENNISON, C. and TURNBULL, B. W. (1991). Exact calculations for sequential t, χ^2, and F tests. *Biometrika* **78** 131–142.

O'BRIEN, P. C. and FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35** 549–556.

SELLKE, T. and SIEGMUND, D. (1983). Sequential analysis of the proportional hazards model, *Biometrika* **70** 315–326.

SIEGMUND, D. (1980). Sequential χ^2 and F tests and the related confidence intervals. *Biometrika* **67** 389–402.

SIEGMUND, D. (1985). *Sequential Analysis: Tests and Confidence Intervals*. Springer, New York.

SIEGMUND, D. (1986). Boundary crossing probabilities and statistical applications. *Ann. Statist.* **14** 361–404.

SIEGMUND, D. (1988). Confidence sets in change-point problems. *Internat. Statist. Rev.* **56** 31–48.

WOODROOFE, M. (1978). Large deviations of the likelihood ratio statistic with applications to sequential testing. *Ann. Statist.* **6** 72–84.

WOODROOFE, M. (1986). Very weak expansions for sequential confidence levels. *Ann. Statist.* **14** 1049–1067.

WOODROOFE, M. (1989). Very weak expansions for sequentially designed experiments: Linear models. *Ann. Statist.* **17** 1087–1102.

DEPARTMENT OF STATISTICS
SEQUOIA HALL
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305