View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Knowledge at Singapore Management University

Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Economics | School of Economics

5-2003

Early stopping by using stochastic curtailment in a three-arm sequential trial

Denis H. Y. LEUNG
Singapore Management University, denisleung@smu.edu.sg

You-Gan WANG
National University of Singapore

David AMAR
Memorial Sloan Kettering Cancer Center

DOI: https://doi.org/10.1111/1467-9876.00394

Follow this and additional works at: https://ink.library.smu.edu.sg/soe_research

Part of the Econometrics Commons

Citation

LEUNG, Denis H. Y.; WANG, You-Gan; and AMAR, David. Early stopping by using stochastic curtailment in a three-arm sequential trial. (2003). *Journal of the Royal Statistical Society - Series C: Applied Statistics*. 52, (2), 139-152. Research Collection School Of Economics.
Available at: https://ink.library.smu.edu.sg/soe_research/106

This Journal Article is brought to you for free and open access by the School of Economics at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Economics by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Published in Journal of the Royal Statistical Society: Series C: Applied Statistics
Volume 52, Issue 2, May 2003, Pages 139-152
https://doi.org/10.1111/1467-9876.00394

Early stopping by using stochastic curtailment in a three-arm sequential trial

Denis Heng-Yan Leung,

Singapore Management University, Singapore, and Memorial Sloan–Kettering Cancer Center, New York, USA

You-Gan Wang

National University of Singapore, Singapore

and David Amar

Memorial Sloan–Kettering Cancer Center, New York, USA

Summary. Interim analysis is important in a large clinical trial for ethical and cost considerations. Sometimes, an interim analysis needs to be performed at an earlier than planned time point. In that case, methods using stochastic curtailment are useful in examining the data for early stopping while controlling the inflation of type I and type II errors. We consider a three-arm randomized study of treatments to reduce perioperative blood loss following major surgery. Owing to slow accrual, an unplanned interim analysis was required by the study team to determine whether the study should be continued. We distinguish two different cases: when all treatments are under direct comparison and when one of the treatments is a control. We used simulations to study the operating characteristics of five different stochastic curtailment methods. We also considered the influence of timing of the interim analyses on the type I error and power of the test. We found that the type I error and power between the different methods can be quite different. The analysis for the perioperative blood loss trial was carried out at approximately a quarter of the planned sample size. We found that there is little evidence that the active treatments are better than a placebo and recommended closure of the trial.

Keywords: Bonferroni adjustment; Conditional power; Interim analysis; Predictive power; Stochastic curtailment; Stopping time

1. Introduction

For ethical and practical reasons sequential designs are commonly used in clinical trials. In sequential trials, we have the option to stop the trial early if data accumulated in the trial strongly suggest the conclusion in favour of one of the treatments. Early stopping in favour of hypothesis H_1 has been well studied (Jennison and Turnbull, 2000). But, sometimes, early stopping in view of a negative result is also desirable (see, for example DeMets and Ware (1980, 1982) and Pepe and Anderson (1992)). In this paper, we consider the opportunity of early stopping for both hypothesis H_0 and hypothesis H_1 in a three-arm randomized double-blind study involving a placebo and two active treatments.

Address for correspondence: Denis Heng-Yan Leung, Singapore Management University, Tanglin PO Box 257, 912409 Singapore.
E-mail: denisleung@smu.edu.sg

Major orthopaedic procedures can be associated with substantial perioperative blood loss (PBL) requiring the transfusion of multiple units of red blood cells. Blood loss may be associated with an increased risk of infection post operatively and longer lengths of hospital stay. Currently, a proven technique to minimize blood loss in this patient population is deliberate hypotension (Thompson *et al.*, 1978). The administration of antifibrinolytic agents has also shown promise in further reducing perioperative bleeding (Vander Salm *et al.*, 1988). To determine the benefits of these antifibrinolytic agents better, a randomized double-blind study was initiated in 1999 at the Memorial Sloan–Kettering Cancer Center to compare two antifibrinolytic agents, e-aminocaproic acid (EACA) and aprotinin with a placebo, in surgical patients at high risk of a significant loss of blood. The study was designed with a maximum sample size of 105 per arm to detect a difference of 30% of blood loss between the arms with a power of 80% and an overall two-sided type I error of 5%.

When the study was designed, it was expected that the accrual rate of these patients would be approximately 160 per year and, therefore, the trial could be completed in no more than 2 years. However, as the trial progressed, it became evident that the accrual could not meet the expected rate—69 patients were accrued to the trial over a span of 18 months. This fact, compounded with emerging data from smaller studies, prompted us to carry out an unplanned interim analysis. With results available for 24, 24 and 22 patients in the placebo, aprotinin and EACA treatment arms, the mean (with standard deviation in parentheses) operative blood loss (on a natural logarithmic scale) results are 6.6217 (0.7886), 6.8167 (1.1929) and 6.7936 (0.8888) respectively. These results indicate that, when considering stopping the trial, the possibility of doing so not only in favour of hypothesis H_1 but also of hypothesis H_0 should be studied.

Many researchers have suggested designs that allow unplanned interim analyses. In particular, Lan *et al.* (1982) suggested a method of stochastic curtailment. In this method, early stopping is based on calculating the conditional power, i.e. the chance that the results at the end of the trial will be significant, given the current data. The method does not restrict the time at which an interim analysis is to be carried out. It is therefore very attractive in practice because interim analyses are often carried out at only approximately equal intervals. Furthermore, unplanned interim analyses can also be accommodated under this paradigm. Similar procedures have been considered by Jennison and Turnbull (1990), Pepe and Anderson (1992) and Betensky (1997a, b).

Other stochastic curtailment methods have been reported. Some researchers considered a *predictive power approach* that involves averaging the conditional power over the posterior of the treatment effects parameter (Herson, 1979; Spiegelhalter *et al.*, 1986). Others considered procedures based on a *conditional probability ratio* (Jennison, 1992; Xiong, 1995). This approach uses a likelihood ratio test of whether the test statistic at the end of the trial will be consistent with the accumulated data, i.e. a test of whether the final analysis (at the maximum sample size) will end in favour of hypothesis H_0 or hypothesis H_1, on the basis of the current data. Another approach is to calculate a *prediction limit* of the test statistic at the end of the trial on the basis of the current data. Finally, the conditional power evaluated at H_0 can also be used. Using this quantity, a small value indicates that the current data are in favour of the null hypothesis whereas a high value supports the alternative.

In this paper, we consider using stochastic curtailment methods to analyse the results from a three-arm randomized study that allows for early stopping in favour of hypothesis H_0 or hypothesis H_1. Several problems in a multiarm study are not encountered in a two-arm study. Specifically, the multiplicity of tests will influence the size and power, and hence complicate the design issues. The question is 'How are the size and power affected when stochastic curtailment is applied?'. Furthermore, the definition of an alternative hypothesis is less clear in a multiarm

study (see, for example, Siegmund (1993)). The question is 'How is the power affected under the various alternatives, when stochastic curtailment is used?'. Finally, we would like to find out the relationship between timing of the curtailment and operating characteristics.

2. Sequential stopping by using stochastic curtailment

In this section, we consider a few stochastic curtailment methods for early stopping under the scenario of a comparison between two treatments. We shall discuss how it extends to a three-arm study in Section 3.

Without loss of generality, we write the two-sided hypothesis testing problem as

$$H_0 : \theta = 0 \textit{ versus } H_1 : \theta = \theta_1 \neq 0,$$

where θ represents the treatment difference between the two arms. Suppose that we have a K-stage sequential test. Let $n_1(k)$ and $n_2(k)$ denote respectively the sample size in the two arms under comparison at the kth stage. This formulation suggests the possibility of different numbers of observations between interim analyses. Also let the means and standard deviations of the treatment effects, after k stages, in the two arms be $\bar{X}_{1(k)}, \bar{X}_{2(k)}$, $s_{1(k)}$ and $s_{2(k)}$. Then the hypotheses can be tested by using the normalized test statistic

$$Z_{n(k)} = \frac{\bar{X}_{1(k)} - \bar{X}_{2(k)}}{\sqrt{\{s^2_{1(k)}/n_1(k) + s^2_{2(k)}/n_2(k)\}}}, \tag{1}$$

where $n(k) = n_1(k) + n_2(k)$. It is essentially the recommendation of Lan and Wittes (1988) in applying a one-arm set-up to the results of a two-arm study. Note that $Z_{n(k)}$ can be considered as a test statistic with information level $n(k)$. We use $m = n(K)$ to denote the information at the last stage, stage K. We assume that $Z_{n(k)}$ is normally distributed with unknown mean θ and known variance. For ease of exposition, we assume that the variance is 1.

The method of stochastic curtailment uses a 'reference test', which can be a fixed sample or a sequential test. For simplicity in the discussion in this section, we assume that the reference test is a fixed sample test with type I error α and power $1 - \beta$ to detect a value of $\theta = \theta_1$. Let b be the critical value for Z_m, i.e. reject hypothesis H_0 if $|Z_m| \geqslant bm^{-1/2}$ and accept H_0 otherwise. The value of m is determined to assure a power of $1 - \beta$ in the reference test.

The conditional power is defined as $\mathrm{CP}(\theta^*) = \Pr(|Z_m| \geqslant bm^{-1/2} | Z_{n(k)}, \theta = \theta^*)$. Lan *et al.* (1982) suggested calculating $\mathrm{CP}(\theta_1)$. If it is less than some threshold γ_0, we should stop the trial in favour of hypothesis H_0. In contrast, for early stopping in favour of H_1, they suggested calculating $\mathrm{CP}(0)$ and recommended accepting H_1 if it is greater than some threshold $1 - \gamma_1$. At time $k \leqslant K$, their suggestion leads to the following rule: accept H_1 if

$$|Z_{n(k)}| \geqslant b\{m/n(k)\}^{1/2} + z_{1-\gamma_1}[\{m - n(k)\}/n(k)]^{1/2};$$

accept H_0 if

$$|Z_{n(k)}| \leqslant b\{m/n(k)\}^{1/2} - \theta_1\{m - n(k)\}\, n(k)^{-1/2} - z_{1-\gamma_0}[\{m - n(k)\}/n(k)]^{1/2}.$$

Here $z_{1-\gamma_0}$ is the $1 - \gamma_0$ normal deviate and $z_{1-\gamma_1}$ is the $1 - \gamma_1$ normal deviate. Therefore, for fixed m the stopping boundaries can also be parameterized by the three parameters (b, γ_0, γ_1). Variations of the method of Lan *et al.* (1982) have been considered by Jennison and Turnbull (1990), who used $\mathrm{CP}(\hat{\theta})$, where $\hat{\theta}$ is an estimate of θ based on the current data. Pepe and Anderson (1992) suggested a small $\mathrm{CP}\{\hat{\theta} + \mathrm{se}(\hat{\theta})\}$ as an indication in favour of H_0. An alternative to the conditional power approach is the so-called predictive power (PP) approach (Jennison and Turnbull, 1990). The PP is defined as the conditional power integrated over the posterior

distribution π of the unknown θ, given the data. If a non-informative prior is used, then the PP is independent of any parameters. For normal data with information level $n(k)$, the posterior of θ is $N\{Z_{n(k)}\, n(k)^{-1/2}, 1/n(k)\}$; therefore accept H_1 if

$$|Z_{n(k)}| \geqslant b\{n(k)/m\}^{1/2} + z_{1-\gamma_1}[\{m - n(k)\}/m]^{1/2}$$

but accept H_0 if

$$|Z_{n(k)}| \leqslant b\{n(k)/m\}^{1/2} - z_{1-\gamma_0}[\{m - n(k)\}/m]^{1/2}.$$

This family of stopping times may also be derived by using the conditional probability ratio (Jennison, 1992; Xiong, 1995).

The conditional power approach has been criticized for its arbitrariness because it is evaluated on the basis of the value of θ under hypothesis H_1 (see, for example, Betensky (2000)). This leads to suggestions such as those in Jennison and Turnbull (1990). Note that, using stochastic curtailment, data monitoring is based on the accumulated data. The projection of the behaviour of future data is a way to assess the possible outcomes of the trial if it is to be continued. As such, any reasonable assumption of the distribution for future observations may be appropriate in establishing an index for the current state of the trial. Therefore, it is also possible to consider a version of conditional power evaluated under H_0. This will alleviate the arbitrariness concerns. In other words, we use

$$c(Z_{n(k)}) = \Pr(|Z_m| \geqslant bm^{-1/2} | Z_{n(k)}, \theta = 0)$$

as a curtailment tool. $c(Z_{n(k)})$ measures how far the current observations deviate from the null hypothesis. At each time k, $1 \leqslant k \leqslant K$, $c(Z_{n(k)})$ is a random variable ranging from 0 to 1 and having a mean of $\alpha = 1 - \Phi(bm^{-1/2})$ because

$$\begin{aligned} E\{c(Z_{n(k)})\} &= E[E\{I(|Z_m| \geqslant bm^{-1/2} | Z_{n(k)}, \theta = 0)\}] \\ &= E\{I(|Z_m| \geqslant bm^{-1/2} | \theta = 0)\} = \alpha. \end{aligned}$$

At the last stage and Kth stage, $c(Z_m)$ follows a Bernoulli distribution with probability $1-\alpha$ at 0 and α at 1. In general, for $1 \leqslant k < K$, the conditional error rate $c(Z_{n(k)})$ is a monotonic function in $Z_{n(k)}$ and it measures how far the current data are against H_0. It is therefore reasonable to argue that, if this error rate falls below some threshold value γ_0, continuation of the trial is unlikely to lead to termination in favour of H_1, whereas, if it rises above some threshold $1 - \gamma_1$, then the trial should stop to reject H_0. This corresponds to stopping in favour of H_1 if

$$\Pr(|Z_m| \geqslant bm^{-1/2} | Z_{n(k)}, \theta = 0) > 1 - \gamma_1$$

and stopping in favour of H_0 if

$$\Pr(|Z_m| \geqslant bm^{-1/2} | Z_{n(k)}, \theta = 0) < \gamma_0,$$

which leads to the following early stopping rules at time $k \leqslant K$: accept H_1 if

$$|Z_{n(k)}| \geqslant b\{m/n(k)\}^{1/2} + z_{1-\gamma_1}[\{m - n(k)\}/n(k)]^{1/2};$$

accept H_0 if

$$|Z_{n(k)}| \leqslant b\{m/n(k)\}^{1/2} - z_{1-\gamma_0}[\{m - n(k)\}/n(k)]^{1/2}.$$

The type I error for the sequential test is $\Pr(\text{stopping at } K, |Z_m| \geqslant bm^{-1/2} | \theta = 0)$, which is always larger than α (when $K \geqslant 2$). Also, the upper boundary is identical with that of early stopping using conditional power (Lan *et al.*, 1982). This leads to the stopping boundaries for

$Z_{n(k)}$ with three design parameters (b, γ_0, γ_1) for a fixed m. In practice, we may choose (b, γ_0, γ_1) to control for the type I error and power (or the expected sample size).

3. Comparison of the methods in a multiarm study

The methods discussed in Section 2 have been used for decision-making in the PBL trial. One of the first problems that we needed to solve before using these methods for decision-making was to study their operating characteristics. Although these methods have been widely used for decision-making in trials involving two treatments, little is known about their characteristics when applied in a multiarm study. In this section, we present some results regarding their operating characteristics in the context of a three-arm study.

Sequential monitoring of multiarm trials has been studied by Hughes (1993), Siegmund (1993), Follmann *et al.* (1994) and Betensky (1996). Hughes (1993) considered the case where all treatments are under direct comparison (pairwise comparison) and the case where one of the treatments is a control. He suggested that, when one treatment is a control, the control should be kept throughout the trial whereas active treatments can be dropped early. Siegmund (1993) studied a two-stage approach whereby the trial proceeds until at least one treatment has been eliminated; then a comparison between the remaining treatments will be made. As in Siegmund (1993), Betensky (1996) also studied a two-stage approach. But, in addition to pairwise comparisons, she also considered the case of a control treatment. Follmann *et al.* (1994) considered the possibility of early stopping in favour of hypothesis H_0. Their method is based on pairwise comparisons between the arms. The type I error adjustment is a generalization of Dunnett's or Tukey's procedure for multiple comparisons. They showed that the adjustments that they proposed produce stopping boundaries which are essentially identical with those by using a simple Bonferroni adjustment in the type I error (Follmann *et al.* (1994), Tables 1 and 2). They subsequently suggested using Bonferroni adjustment because it is simpler and more flexible, for example, when the amount of information is different between each pair of arms. Furthermore, it is a widely established method for effectively controlling the type I error when the number of groups is not too large. Therefore, in this section, we consider a trial designed with a Bonferroni adjustment to the type I error.

Unlike a trial with two arms, where the hypothesis to be tested is straightforward, namely no difference *versus* some differences between the arms, in a three-arm study there are several possible ways to formulate the problem.

In the PBL trial, two active treatments (aprotinin and EACA) have been compared with a single control. In that situation, Hughes (1993) suggested that the control arm should not be dropped in any of the interim analyses whereas other arms may be dropped if they are found to be inferior to any other arms. Let T_1 and T_2 be the active treatments and C be the control: the interesting situations are

(a) $T_1 = T_2 = C$ (the null hypothesis)—the elimination of any treatment leads to a type I error,
(b) $T_1 > T_2 = C$—a failure to eliminate T_2 or C (at the last look) or the elimination of T_1 both lead to errors,
(c) $T_1 = T_2 > C$—the elimination of T_1 or T_2 or a failure to eliminate C (at the last look) both lead to errors—and
(d) $T_1 > T_2 > C$—a failure to choose T_1 as the superior treatment results in an error.

There are also situations where all three arms are active. We consider these scenarios for completeness. Let T_1, T_2 and T_3 denote the treatment effects of the treatments under study. Then

Siegmund (1993) identified some situations that are interesting to study:

(a′) $T_1 = T_2 = T_3$ (the null hypothesis)—the elimination of any treatment leads to a type I error,
(b′) $T_1 > T_2 = T_3$—a failure to eliminate T_2 or T_3 or the elimination of T_1 both lead to errors,
(c′) $T_1 = T_2 > T_3$—the elimination of T_1 or T_2 or a failure to eliminate T_3 both lead to errors—and
(d′) $T_1 > T_2 > T_3$—a failure to choose T_1 as the superior treatment results in an error.

We studied the behaviours of five stochastic curtailment methods when

(i) all arms are under direct comparison and
(ii) one of the arms is a control.

In each case, we assumed that the reference test is a fixed sample two-sided test for

$$H_0^* : T_i - T_j = 0 \textit{ versus } H_1^* : T_i - T_j = \theta_1 \neq 0, i \neq j,$$

with overall type I error of 5%, following a Bonferroni adjustment and an 80% power to detect θ_1. The five methods are CP_1 (Lan *et al.*, 1982), CP_2 (Jennison and Turnbull, 1990), CP_3 (Pepe and Anderson, 1992), PP and CP_4 (the conditional power function but evaluated at $\theta = 0$, i.e. $c(Z_m)$). The CP_3 method is a modification of Pepe and Anderson's method (which only considered early stopping in favour of hypothesis H_0) to allow also early stopping in favour of H_1 if $CP\{\hat{\theta} - se(\hat{\theta})\} > 1 - \gamma_1$. For each method, the value of b is that which is used by the corresponding reference test. The value of γ_0 is constrained to be the same as γ_1 for each method.

We first studied the effect of using stochastic curtailment methods on the overall type I error and power. We considered sequential trials with $K = 2$ and $K = 5$ looks. For each method, we used three values of $\gamma_0 = \gamma_1 = \gamma$ that correspond to threshold values of 0.001, 0.05 and 0.2. We constrained the maximum number of observations in the sequential trial to be the same as that in the reference test. Furthermore, any test of the sample size m can be standardized to a test of sample size 1. For example, a test based on Z_m can be considered to be a test with the sample size of the reference test normalized to be the same as the number of looks, i.e. we considered a reference test with sample sizes of 2 and 5.

We defined power as the complement of the error rate, i.e. 1 minus the error rate, where the error is as defined under situations (b), (c) and (d) or under (b′), (c′) and (d′). Furthermore, under situations (b), (c) and (d) and (b′), (c′) and (d′), differences between treatments are fixed at θ_1. For example, in case (c′), the power of the test is the probability of not rejecting hypothesis H_0 when $T_1 = T_2 = T_3 + \theta_1$, but in case (d′) the power of the test is the probability of not rejecting H_0 when $T_1 = T_2 + \theta_1$; $T_2 = T_3 + \theta_1$. In this study, we fixed the value of θ_1 at 2.28 and 1.44 respectively for the two- and five-looks situation. These values of θ_1 were chosen such that there is an 80% chance that they can be detected by using a fixed sample size test (reference) with a sample size of 2 and 5 respectively. This set-up is completely general so, for example, the PBL trial was planned for a maximum of 210 patients for comparison between any two arms. In that case, the fixed sample size test is based on $m = 210$, instead of $m = 2$, observations. The tabulated results then correspond to a test with an 80% chance of detecting a value of $\theta_1 = 2.28(m/2)^{-1/2} = 2.28(105)^{-1/2} = 0.222$.

All the methods considered use b, which corresponds to the critical value for a fixed sample size test based on m observations, i.e. reject hypothesis H_0 if $|Z_m| \geqslant bm^{-1/2}$ and accept H_0 otherwise. The value of b determines the type I error. In the simulations, the type I error of the fixed sample test is limited to less than 5% after Bonferroni adjustment. So the value of b is $z_{\alpha/2}$

where $\alpha = 0.05/(\text{number of pairwise comparisons})$, and $z_{1-\alpha}$ is the $(1-\alpha/2)\%$ normal deviate. For three arms, the number of comparisons is 3, so $b = 2.39\alpha/2$.

For each situation that we studied, 200 000 simulations were used to estimate the type I error and power. The results when all the arms are under direct comparison are given in Table 1 and when one arm is a control are given in Table 2. We also calculated the Monte Carlo average sample number ASN under hypothesis H_0 for each method.

Tables 1 and 2 record results based on $K = 2$ and $K = 5$ looks. The patterns of results are very similar for each of these two scenarios. Therefore, in discussing the results in Tables 1 and 2, we concentrate on the two-looks case and refer the readers to Tables 1 and 2 for results under $K = 5$. In Table 1, the behaviours of the various methods are very similar when the value of γ is small, i.e., when early stopping in favour of hypothesis H_0 or H_1 is based on a very stringent threshold, the behaviours of the methods are very similar. Overall, the type I error

Table 1. Type I error, power and average sample number ASN under hypothesis H_0 by using five stochastic curtailment methods in a three-arm study (direct comparison)†

Number of looks	$\gamma_0 = \gamma_1$	*Method*	*Type I error*	*Power*‡	*Power*§	*Power*§§	*ASN*
2	0.001	CP_1	0.044	0.69	0.79	0.80	1.99
		CP_2	0.046	0.69	0.79	0.80	1.98
		CP_3	0.044	0.69	0.79	0.80	1.99
		PP	0.044	0.69	0.79	0.80	1.99
		CP_4	0.044	0.69	0.78	0.80	1.95
	0.050	CP_1	0.044	0.69	0.79	0.80	1.99
		CP_2	0.064	0.70	0.79	0.81	1.64
		CP_3	0.045	0.69	0.80	0.80	1.99
		PP	0.051	0.69	0.79	0.80	1.85
		CP_4	0.029	0.64	0.60	0.80	1.19
	0.200	CP_1	0.044	0.69	0.78	0.80	1.96
		CP_2	0.106	0.72	0.75	0.83	1.37
		CP_3	0.052	0.69	0.79	0.81	1.81
		PP	0.083	0.71	0.77	0.82	1.49
		CP_4	0.010	0.47	0.35	0.79	1.03
5	0.001	CP_1	0.044	0.69	0.79	0.80	4.91
		CP_2	0.048	0.69	0.80	0.80	4.52
		CP_3	0.044	0.69	0.80	0.80	4.84
		PP	0.044	0.69	0.80	0.80	4.72
		CP_4	0.043	0.69	0.77	0.80	4.28
	0.050	CP_1	0.044	0.69	0.79	0.80	4.48
		CP_2	0.233	0.71	0.68	0.84	2.89
		CP_3	0.045	0.69	0.80	0.80	4.33
		PP	0.056	0.70	0.80	0.81	3.88
		CP_4	0.007	0.24	0.18	0.40	1.18
	0.200	CP_1	0.043	0.69	0.77	0.80	4.00
		CP_2	0.378	0.71	0.57	0.86	2.16
		CP_3	0.059	0.70	0.79	0.81	3.74
		PP	0.223	0.71	0.68	0.85	2.76
		CP_4	<0.001	0.01	0.01	0.03	1.00

†Each method uses the same b-value as a fixed sample test with overall two-sided type I error 0.05 and power 80%, with a Bonferroni adjustment. Entries were obtained by 200 000 simulations.
‡$T_1 > T_2 = T_3$.
§$T_1 = T_2 > T_3$.
§§$T_1 > T_2 > T_3$.

Table 2. Type I error, power and average sample number ASN under hypothesis H_0 by using five stochastic curtailment methods in a three-arm study (treatments *versus* control)†

Number of looks	$\gamma_0 = \gamma_1$	*Method*	*Type I error*	*Power*‡	*Power*§	*Power*§§	*ASN*
2	0.001	CP_1	0.044	0.69	0.68	0.80	1.99
		CP_2	0.045	0.69	0.68	0.80	1.98
		CP_3	0.044	0.69	0.68	0.80	1.99
		PP	0.044	0.69	0.68	0.80	1.99
		CP_4	0.044	0.69	0.68	0.80	1.94
	0.050	CP_1	0.044	0.69	0.68	0.80	1.99
		CP_2	0.056	0.69	0.67	0.81	1.58
		CP_3	0.045	0.69	0.68	0.80	1.99
		PP	0.048	0.69	0.68	0.80	1.81
		CP_4	0.024	0.58	0.64	0.80	1.15
	0.200	CP_1	0.044	0.69	0.68	0.80	1.95
		CP_2	0.082	0.68	0.65	0.83	1.31
		CP_3	0.049	0.69	0.68	0.81	1.77
		PP	0.067	0.68	0.67	0.82	1.42
		CP_4	0.007	0.35	0.47	0.79	1.02
5	0.001	CP_1	0.044	0.69	0.68	0.80	4.88
		CP_2	0.046	0.69	0.68	0.80	4.41
		CP_3	0.044	0.69	0.68	0.80	4.79
		PP	0.044	0.69	0.68	0.80	4.66
		CP_4	0.042	0.69	0.68	0.80	4.16
	0.050	CP_1	0.042	0.69	0.68	0.80	4.40
		CP_2	0.173	0.66	0.60	0.84	2.70
		CP_3	0.042	0.69	0.68	0.80	4.23
		PP	0.048	0.69	0.68	0.81	3.71
		CP_4	0.005	0.21	0.31	0.66	1.13
	0.200	CP_1	0.038	0.68	0.68	0.80	3.88
		CP_2	0.281	0.64	0.51	0.87	2.11
		CP_3	0.049	0.68	0.68	0.81	3.59
		PP	0.164	0.67	0.61	0.85	2.61
		CP_4	<0.001	0.01	0.02	0.19	1.00

†Each method uses the same b-value as a fixed sample test with overall two-sided type I error 0.05 and power 80%, with a Bonferroni adjustment. Entries were obtained by 200 000 simulations.
‡$T_1 > T_2 = C$.
§$T_1 = T_2 > C$.
§§$T_1 > T_2 > C$.

is close to the desired 5% level and the power is uniformly high for each method. This is not surprising because the chance of early stopping is small and therefore the behaviours should be very similar to a fixed sample test applied to multiple comparisons. This also results in an ASN that is close to the number of looks (for example, $\gamma = 0.001$ and ASN = 1.99 for CP_1). As γ increases, the chance of curtailment in either direction increases. However, it is seen that for the CP_1 and CP_3 methods the overall type I error and power under all situations are very close to the case when γ is small. In contrast, the type I errors of the PP and CP_2 methods increase to 0.083 and 0.106 respectively for $\gamma = 0.2$, with corresponding drops in power. The power for the CP_4 method drops substantially for larger values of γ. These results are not surprising. CP_4 is a conditional probability under H_0. Therefore, even a moderate value of γ leads to very liberal stopping in favour of H_0, giving a very small type I error but poor power. This, however, is not an indication that CP_2, CP_4 or PP are poor methods. It only suggests that the values of b, γ_0 and γ_1 must be carefully chosen for these methods. Note also that the power can exceed 80% because in situation (d) the underlying difference between the best treatment T_1 and T_3 is

$\theta_1 + \theta_1$, which is bigger than the difference of θ_1 assumed in the calculation of the fixed sample reference test.

When one arm is a control (Table 2), the power is relatively stable for all the methods except CP_4 where the power drops precipitously as the value of γ increases. For example, CP_4's power to detect $T_1 = T_2 > C$ is 0.69 when $\gamma = 0.001$ is used. But its power for the same test drops to 0.58 and 0.35 when $\gamma = 0.05$ and $\gamma = 0.2$ respectively are used. This result once again underscores the fact that the interpretation of γ is different for the different methods. The sizes for the PP and CP_2 methods once again are inflated for larger values of γ (0.067 and 0.082 for $\gamma = 0.2$). However, the size inflations for these methods are smaller than the case when all treatments are under direct comparison. This can be explained by the fact that early dropping of the control is not possible. Once again, method CP_4 is anticonservative for larger values of γ.

So far, we have considered curtailments performed at one of the planned interim analyses. However, one of the most attractive properties of curtailment is that the analysis can occur at

Table 3. Effect of timing of the interim analysis on the type I error, power and average sample number ASN under hypothesis H_0 in a three-arm study (direct comparison)†

f	$\gamma_0 = \gamma_1$	*Method*	*Type I error*	*Power*‡	*Power*§	*Power*§§	*ASN*
0.25	0.001	CP_1	0.044	0.69	0.79	0.80	1.99
		CP_2	0.044	0.69	0.79	0.80	1.99
		CP_3	0.044	0.69	0.79	0.80	1.99
		PP	0.044	0.69	0.79	0.80	1.99
		CP_4	0.044	0.69	0.79	0.80	1.99
	0.050	CP_1	0.044	0.69	0.79	0.80	1.99
		CP_2	0.164	0.71	0.74	0.83	1.81
		CP_3	0.044	0.69	0.79	0.80	1.99
		PP	0.044	0.69	0.79	0.80	1.99
		CP_4	0.015	0.43	0.32	0.75	1.13
	0.200	CP_1	0.044	0.69	0.79	0.80	1.99
		CP_2	0.279	0.72	0.65	0.85	1.54
		CP_3	0.044	0.69	0.79	0.80	1.99
		PP	0.160	0.71	0.74	0.83	1.82
		CP_4	0.001	0.07	0.04	0.41	1.01
0.75	0.001	CP_1	0.045	0.69	0.79	0.80	1.99
		CP_2	0.045	0.69	0.80	0.80	1.74
		CP_3	0.045	0.69	0.79	0.80	1.96
		PP	0.045	0.69	0.79	0.80	1.86
		CP_4	0.044	0.69	0.78	0.80	1.59
	0.050	CP_1	0.045	0.69	0.79	0.80	1.65
		CP_2	0.048	0.69	0.80	0.81	1.35
		CP_3	0.045	0.69	0.80	0.80	1.57
		PP	0.046	0.69	0.80	0.80	1.42
		CP_4	0.038	0.68	0.71	0.80	1.17
	0.200	CP_1	0.044	0.69	0.78	0.80	1.37
		CP_2	0.060	0.71	0.79	0.82	1.18
		CP_3	0.048	0.69	0.80	0.81	1.33
		PP	0.056	0.70	0.79	0.81	1.21
		CP_4	0.026	0.66	0.62	0.80	1.06

†Assuming a study with $K = 2$ looks at $f = 0.25$ and $f = 0.75$ of the maximum number of observations. Entries were obtained by 200 000 simulations.
‡$T_1 > T_2 = T_3$.
§$T_1 = T_2 > T_3$.
§§$T_1 > T_2 > T_3$.

any time over the course of the trial. This is the case in the PBL trial, for example, where an interim analysis was performed on the basis of approximately a quarter of the planned maximum sample size. Here, we give results from a study of the influence of timing of the analysis in a three-arm trial. We considered a trial with two looks and the interim analysis occurring when 25% and 75% of the observations have been accrued. The results when all arms are under direct comparison are given in Table 3, and the results when one arm is a control are given in Table 4. For brevity the corresponding results for an analysis at 50% of the observations of those in Tables 1 and 2 are not presented. Table 3 indicates that, for method CP_1, the timing of the interim analysis has little effect on the overall type I error or the power of the method. In certain cases, however, there is some reduction in ASN if the interim analysis is carried out late in the trial. For example, for $\gamma = 0.05$, ASN = 1.99 at $f = 0.25$ but ASN = 1.65 at $f = 0.75$. This reduction is due to the fact that for the CP_1 method stopping is not allowed at $f = 0.25$ but is possible at $f = 0.75$. Similar results are seen for the CP_3 method. The type I error is

Table 4. Effect of timing of the interim analysis on the type I error, power and average sample number ASN under hypothesis H_0 in a three-arm study (treatments *versus* control)†

f	$\gamma_0 = \gamma_1$	*Method*	*Type I error*	*Power*‡	*Power*§	*Power*§§	*ASN*
0.25	0.001	CP_1	0.044	0.69	0.68	0.80	1.99
		CP_2	0.044	0.69	0.68	0.80	1.99
		CP_3	0.044	0.69	0.68	0.80	1.99
		PP	0.044	0.69	0.68	0.80	1.99
		CP_4	0.044	0.69	0.68	0.80	1.99
	0.050	CP_1	0.044	0.69	0.68	0.80	1.99
		CP_2	0.127	0.69	0.64	0.83	1.80
		CP_3	0.044	0.69	0.68	0.80	1.99
		PP	0.044	0.69	0.68	0.80	1.99
		CP_4	0.011	0.33	0.42	0.74	1.10
	0.200	CP_1	0.044	0.69	0.68	0.80	1.99
		CP_2	0.210	0.66	0.57	0.84	1.52
		CP_3	0.044	0.69	0.68	0.80	1.99
		PP	0.124	0.69	0.64	0.83	1.81
		CP_4	0.001	0.04	0.07	0.40	1.00
0.75	0.001	CP_1	0.045	0.69	0.68	0.80	1.99
		CP_2	0.045	0.69	0.68	0.80	1.68
		CP_3	0.045	0.69	0.68	0.80	1.94
		PP	0.045	0.69	0.68	0.80	1.82
		CP_4	0.044	0.69	0.68	0.80	1.51
	0.050	CP_1	0.044	0.69	0.68	0.80	1.58
		CP_2	0.043	0.69	0.68	0.81	1.29
		CP_3	0.044	0.69	0.68	0.80	1.50
		PP	0.043	0.69	0.68	0.80	1.35
		CP_4	0.032	0.66	0.68	0.80	1.12
	0.200	CP_1	0.041	0.68	0.68	0.80	1.30
		CP_2	0.049	0.69	0.68	0.82	1.14
		CP_3	0.043	0.69	0.68	0.81	1.27
		PP	0.047	0.69	0.68	0.81	1.16
		CP_4	0.020	0.61	0.65	0.80	1.04

†Assuming a study with $K = 2$ looks at first look $f = 0.25$ and $f = 0.75$ of the maximum number of observations. Entries were obtained by 200 000 simulations.
‡$T_1 > T_2 = C$.
§$T_1 = T_2 > C$.
§§$T_1 > T_2 > C$.

inflated for the PP (= 0.16) and CP_2 (= 0.28) methods if the first interim analysis is performed too early ($f = 0.25$ and $\gamma = 0.2$). CP_4, in contrast, is conservative if the first interim analysis is performed too early ($f = 0.25$ and $\gamma \geqslant 0.05$). When the first interim analysis is performed with 75% of the planned maximum sample size, all methods, except CP_4, give the appropriate size; the CP_4 method is too conservative in that case.

The situation is similar when one of the arms is a control (Table 4), i.e. the CP_1 and CP_3 methods are little affected by the timing of the analysis; the PP, CP_2 and CP_4 methods are affected if the first interim analysis is performed too early. Note that the type I errors of all the methods are smaller than their corresponding entries in Table 3 because the control group cannot be dropped early.

4. Results of the prophylactic trial

We now return to the three-arm PBL trial. With results available for 24, 24 and 22 patients for the placebo, aprotinin and EACA arms, the mean (with standard deviation in parentheses) of the operative blood loss (on a natural log-scale) results are 6.6217 (0.7886), 6.8167 (1.1929) and 6.7936 (0.8888) respectively. Normal probability plots (which are not shown) show no evidence of a departure from normality, in each of the three groups. The test statistic (1) is used for comparison. The study was designed to have a maximum of $K = 2$ looks with a maximum sample of 105 per arm or 210 between two arms. So the comparison between the placebo and EACA is based on $46/210 \times 100\% = 21.9\%$ of the maximum sample size. Similarly, between the placebo and aprotinin, and between aprotinin and EACA, the percentages are both 22.9%.

The study in Section 3 shows that stopping early to accept hypothesis H_0 on the basis of few observations is not possible for a few of the methods (Table 4, $f = 0.25$). Indeed, when we used $\gamma = 0.2$, only the CP_2 and CP_4 methods favour stopping to accept no difference between the placebo and the two active treatments; for the other methods, stopping in favour of no difference is not possible by design. However, we were concerned with the unpromising results that were seen in the active treatments. We therefore calculated the probabilities that the active treatments will be shown to be better than the placebo, if the trial were allowed to carry on. These results are given in Table 5.

We shall first focus on the comparison between the control (placebo) arm and the treatment arms. Note that all probabilities are directional. For example, comparing the placebo with aprotinin, the conditional power (CP_1 method) is 0.6843 for placebo better than aprotinin, but

Table 5. Results of the three-arm prophylactic trial: Z-values and powers by using different stochastic curtailment methods

Comparison	$Z_{n(1)}$	*Powers for the following methods:*				
		PP	CP_1	CP_2	CP_3	CP_4
Aprotinin better than placebo	−0.6680	0.0195	0.4023	<0.0001	0.0066	0.0010
Placebo better than aprotinin	0.6680	0.2937	0.6843	0.1282	0.7588	0.0091
EACA better than placebo	−0.6913	0.0202	0.4146	<0.0001	0.0063	0.0011
Placebo better than EACA	0.6913	0.3136	0.6972	0.1498	0.8025	0.0096
EACA better than aprotinin	−0.0749	0.0881	0.5441	0.0019	0.1582	0.0030
Aprotinin better than EACA	0.0749	0.1184	0.5753	0.0057	0.2611	0.0038

it is 0.4023 for aprotinin better than placebo. The primary focus here is whether there is sufficient evidence that the two active treatments are better than the placebo to be worthy of further accrual. Except for CP_1, all the other methods show little evidence of this. For example, the value of PP is only 0.0195 for aprotinin better than placebo and 0.0202 for EACA better than placebo. These values suggest that, even if the trial were to continue, there will only be about a 2% chance that a significant result will be seen in favour of either active treatments. Ethically, we would have great difficulty in continuing a trial on the basis of these numbers. Similar results are seen for the comparison between EACA and the placebo; all the methods, except CP_1, give a low probability that the trial will eventually end in favour of EACA.

CP_1 is the only method that gives support to continuing the trial. This is because, with only a quarter of the observations accrued and with a hypothesized treatment effect of θ_1 for future observations, there is still a considerable chance that the trial will end in favour of one of the arms. This study underscores how important the value of θ_1 can be for decision-making when method CP_1 is used. The trial was designed with a hypothesized value of θ_1 that is much larger than the treatment effects that were seen at the interim analysis. If we assume that the hypothesized value of θ_1 is still plausible, much more data are needed to reject the alternative hypotheses (that at least one of the active arms is better than the placebo). However, we must also balance this assumption with benefits to patients. So the question that an investigator must ask is the plausibility of the hypothesized value of θ_1. It is not uncommon that θ_1 represents the smallest treatment effects of clinical significance that we *wish* to detect. In this regard, its value bears little relationship to the *true* treatment effects. Therefore, it is a leap of faith to assume that such treatment effects will be seen in future data, when the current data do not support such a value.

The results in Table 5 also illustrate the relationship between the different boundaries. That for method CP_3 is always higher than that for method CP_2 because it is conditioned on a treatment difference of $\hat{\theta} + \text{se}(\hat{\theta})$, instead of $\hat{\theta}$ (as in method CP_2) for future observations. The PP is a conditional power averaged over the posterior of θ. Since the posterior is $N\{\hat{\theta}, 1/n(k)\}$, the PP tends to give less weight to the observed treatment effects in decision-making. For example, in testing whether aprotinin is better than the placebo (Table 5, first row), the observed treatment effects statistic is $Z_{n(1)} = -0.668$, a value that does not support the hypothesis that aprotinin is better. In that case, the PP is higher than that for the CP_2 or CP_3 methods. So compared with CP_2 and CP_3 the PP projects a higher chance of a significant result in favour of aprotinin. However, when the observed treatment difference clearly favours the hypothesis (Table 5, second row), the PP can be lower than the powers for the CP_2 or CP_3 methods. In either of these cases, the observed treatment effects become diffused by the posterior.

The CP_4 method gives the conditional power assuming that $\theta = 0$. Its relationship to the other conditional power approaches is therefore dependent on the relationships of θ_1, $\hat{\theta}$ and $\hat{\theta} + \text{se}(\hat{\theta})$ to 0. For example, in the comparison between aprotinin and placebo (Table 5, first row), $\hat{\theta} = -0.195$ and $\hat{\theta} + \text{se}(\hat{\theta}) = -0.195 + 0.292 = 0.097$, so the value for the CP_4 method is higher than that for CP_2 but it is lower than for CP_1 or CP_3.

To date, only one other trial in which patients undergoing orthopaedic tumour surgery were randomized to aprotinin ($n = 13$) or placebo ($n = 12$) has been reported. The trial results showed a statistical difference in blood loss in favour of aprotinin. This has not been reproduced nor accepted as a standard of care. In the PBL trial, we nearly doubled the sample size in each of the three groups and yet most of the methods that we employed showed that there is no trend that either active treatments will be shown to be significantly better than the placebo, even if the trial is to continue. Aprotinin and EACA are not benign drugs since they may enhance clotting after surgery and may cause a potentially fatal pulmonary embolism. Thus, it is safer to stop the trial if the drugs are highly unlikely to benefit the patients and can potentially harm them.

5. Discussion

In this paper, we considered stochastic curtailment for the early termination of a multiarm study. We distinguished two scenarios:

(a) when all the treatments are under direct comparison and
(b) when one of the treatments is a control.

We assumed that the reference test is a two-sided reference test with a Bonferroni adjustment to control the overall type I error due to multiple comparisons. All the methods that were considered here use a 'power' function that is evaluated on the basis of the observed data and a hypothesized value of the treatment effects. We found that, when the stochastic curtailment methods use a critical value b that is identical with that used by the reference test, the overall type I error can be quite different depending on the choice of the threshold parameters (γ_0 and γ_1) and the particular method. For small values of γ_0 and γ_1, the overall type I errors of the methods are very close to the nominal level. For larger values of γ_0 and γ_1, the type I errors of the methods are quite different. This highlights the fact that, although all the methods can be used to stop in favour of or against hypothesis H_0, the power function used by the methods has very different interpretations. For example, conditional power (CP_1) gives the probability that the trial will end in favour of H_1 when H_1 is in fact true. Obviously a moderately small value for the CP_1 method would be sufficient to stop in favour of H_0. In contrast, the CP_4 method asks the chance that the trial will end in favour of H_1 when H_0 is true. Using this pessimistic assumption, a much smaller value is needed to convince the investigators to stop the trial. However, this does *not* preclude methods like CP_4 from being used as a curtailment tool. As long as the difference in the interpretation of the different powers is recognized, all the methods can be used. In fact, in a trial with two looks, all the methods in this study can be made equivalent.

We found that the time of the interim analysis can have a significant effect on the overall type I error of the test. To be specific, when the b-value is unadjusted, then the type I error is inflated for the PP and CP_2 methods, if the interim analysis is performed on the basis of only a fraction of the expected maximum sample size in a trial. However, as suggested above, an adjustment of the b-value would give tests with the desired size.

In the PBL trial, we found that the CP_3 and PP methods do not allow stopping (by design) after an accrual of 25% of the expected maximum sample size, yet both methods indicate a low power for concluding superiority of the active treatments if the trial is to carry on. These results may compel us to stop the trial in favour of hypothesis H_0 even though the inner boundaries have not been breached, as it becomes ethically difficult to continue with treatments that have little chance of benefits. (Indeed, combining these results with the evidence from methods CP_2, CP_4 and slow accrual, we decided to stop the trial.) This kind of situation can be avoided if strict rules of *double* blinding are adhered to. This is possible for certain trials if the observation of results does not reveal the identity of the treatment. But, for some trials, this is not possible. As an example, in a recent double-blind placebo-controlled trial to study arrythmias after surgery, once the end point had been reached, the randomization code was broken to allow treatment of the symptoms (Amar *et al.*, 2000).

Finally, as suggested in Section 3, any of the stochastic curtailment methods trace out a set of stopping boundaries and these can be used to design a trial with a specific size and power.

Acknowledgements

Part of this research was carried out while the first author was at Memorial Sloan–Kettering

Cancer Center. We thank the Joint Editor, the Associate Editor and the referees for their insightful, constructive and detailed comments that led to a substantially improved paper.

References

Amar, D., Roistacher, N., Rusch, V., Leung, D., Ginsburg, I., Zhang, H., Bains, M., Downey, R., Korst, R. and Ginsburg, R. (2000) Effects of diltiazem prophylaxis on the incidence and clinical outcome of atrial arrythmias after thoracic surgery. *J. Thor. Cardvasc. Surg.*, **120**, 790–798.

Betensky, R. (1996) An O'Brien-Fleming sequential trial for comparing three treatments. *Ann. Statist.*, **24**, 1765–1791.

Betensky, R. (1997a) Conditional power calculations for early acceptance of H_0 embedded in sequential tests. *Statist. Med.*, **16**, 465–477.

Betensky, R. (1997b) Early stopping to accept H_0 based on conditional power: approximations and comparisons. *Biometrics*, **53**, 794–806.

Betensky, R. (2000) Alternative derivations of a rule for early stopping in favour of H_0. *Am. Statistn*, **54**, 35–39.

DeMets, D. and Ware, J. (1980) Group sequential methods for clinical trials with a one-sided hypothesis. *Biometrika*, **67**, 651–660.

DeMets, D. and Ware, J. (1982) Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika*, **69**, 661–663.

Follmann, D., Proschan, M. and Geller, N. (1994) Monitoring pairwise comparisons in multi-armed clinical trials. *Biometrics*, **50**, 325–336.

Herson, J. (1979) Predictive probability early termination plans for Phase II clinical trials. *Biometrics*, **35**, 775–783.

Hughes, M. D. (1993) Stopping guidelines for clinical trials with multiple treatments. *Statist. Med.*, **12**, 901–915.

Jennison, C. (1992) Bootstrap tests and confidence intervals for a hazard ratio when the number of observed failures is small, with application to group sequential survival studies. In *Computing Science and Statistics*, pp. 89–97. New York: Springer.

Jennison, C. and Turnbull, B. (1990) Statistical approaches to interim monitoring of medical trials: a review and commentary. *Statist. Sci.*, **5**, 299–317.

Jennison, C. and Turnbull, B. (2000) *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton: Chapman and Hall–CRC.

Lan, K., Simon, R. and Halperin, M. (1982) Stochastically curtailed tests in long-term clinical trials. *Communs Statist.*, **1**, 207–219.

Lan, K. and Wittes, J. (1988) The B-value: a tool for monitoring data. *Biometrics*, **44**, 579–585.

Pepe, M. S. and Anderson, G. L. (1992) Two-stage experimental designs: early stopping with a negative result. *Appl. Statist.*, **41**, 181–190.

Siegmund, D. (1993) A sequential clinical trial for comparing three treatments. *Ann. Statist.*, **21**, 464–483.

Spiegelhalter, D., Freedman, L. and Blackburn, P. (1986) Monitoring clinical trials: conditional or predictive power? *Contr. Clin. Trials*, **7**, 8–17.

Thompson, G. E., Miller, R. D., Stevens, W. C. and Murray, W. R. (1978) Hypotensive anesthesia for total hip arthroplasty: a study of blood loss and organ function. *Anesthesiology*, **48**, 91–96.

Vander Salm, T. J., Ansell, J. E., Okike, O. N., Marsicano, T. H., Lew, R., Stephenson, W. P. and Rooney, K. (1988) The role of epsilon-aminocaproic acid in reducing bleeding after cardiac operation: a double-blind randomized study. *J. Thor. Cardvasc. Surg.*, **95**, 538–540.

Xiong, X. (1995) A class of sequential conditional probability ratio tests. *J. Am. Statist. Ass.*, **90**, 1463–1473.