ith sufficient numbers of failures, the intervals based on maximum partial likelihood should be the same as those obtained by inverting the score statistic (4.5). Since almost all software routinely in monitoring studies with survival end points calculates these estimates and their estimated ard errors, these intervals are easily constructed, though they should clearly be interpreted with on at the early stages of a study when few failures have been observed.

ichael D. Hughes (Royal Free Hospital School of Medicine, London): If this paper encourages the of confidence intervals in clinical trials then it should be welcomed. The increased width of the vals derived would help to reduce the 'significance' of the particularly small nominal p values that mphasized when a clinical trial stops early: these p values are meaningless as a measure of the type or rate though they are often interpreted as such. However, it does concern me that the intervals are metric and, for a normal response with known variance, this is about the sample mean observed at th stage (equation (2.9)). Although this mean may be, approximately, unbiased with respect to the gn adopted, it is important to note that it must necessarily be large to facilitate stopping. Hughes and ock (1988) illustrate the dramatic effects that this can have notably if the trial is stopped very early: observed mean then relates more to the stopping boundary used than the underlying true mean. To nter the implausibility of these observed values, some shrinkage towards θ_0, as specified in the null othesis, should be expected.

am glad that the authors recognize that stopping a clinical trial is not primarily a statistical decision. With this in mind, a Bayesian approach gives an opportunity to explore how sensitive the interpretation the result obtained is to different pre-trial (prior) belief distributions, chosen to reflect the views of the investigators, other clinicians, trial financiers and other interested parties. Although such ideas require further work, they will generally produce appropriate shrinkage and would help to emphasize the uncertainty in stopping a trial in the face of what are often unexpectedly surprising results.

Professor T. L. Lai (Stanford University): I have a few minor disagreements with the authors' terminology. The term 'acute responses' in Section 1.2 may be confusing to medical practitioners, especially in the light of acute versus chronic diseases. I prefer to use the term 'immediate responses' instead. The authors say at several places that repeated confidence intervals provide a useful 'summary' of the data. For example, in Section 4.1.4, they say that a repeated confidence interval for the hazard ratio between two treatments is a 'useful summary of survival information' at the termination of a trial. In practice, however, we usually summarize survival data in the form of life-tables or graphs of Kaplan–Meier curves. These provide much more easily understood and interpretable summaries of the data for readers of medical journals than a set of different confidence intervals $I_1, \ldots, I_K$ of the hazard ratio (in a postulated proportional hazards model), which have been computed during several interim analyses. Repeated confidence intervals are for making valid sequential inferences, in a frequentist mode, in the absence of a prespecified stopping rule in the protocol (see Jennison and Turnbull (1984) and Lai (1984)). In this connection, I still prefer the seemingly less specific term 'confidence sequence' to the authors' 'repeated confidence intervals', which may appear to the medical readers, at first sight, to be counterparts in estimation problems of Armitage's 'repeated significance tests'.

In its ordinary ('non-calculus') usage, a 'sequence' need not be infinite. Thus, repeated confidence intervals are simply a finite sequence of confidence intervals with a prescribed probability of simultaneously covering the true parameter. The problem of computing coverage probabilities of confidence sequences (finite or infinite) for univariate or multivariate parameters has been covered substantially in the literature. For the univariate normal case of Section 2.2, the Armitage–McPherson–Rowe recursive numerical integration algorithm can be used to compute the boundary crossing probabilities. The authors rely on using Wiener process approximations to reduce the other problems that they discuss to this special case and have found that these approximations are sometimes unsatisfactory. For some of these problems, more refined approximations have been developed in the recent literature on boundary crossing probabilities: see Siegmund (1985).

Dr D. Y. Lin and Professor L. J. Wei (University of Wisconsin, Madison): We note that the authors generally use the direct normal approximation to the distribution of the maximum likelihood estimator (MLE) to construct repeated confidence intervals (RCIs). Such RCIs are simple to obtain in practice. However, as in the one-stage analysis, these procedures are not always appealing (see Cox and Hinkley (1974), pp. 342–343). Alternatively, using Taylor series expansion for the logarithm of the likelihood ratio statistic (LRS) (see Cox and Hinkley (1974), p. 323), we can obtain approximate RCIs by inverting

TABLE 11

Empirical levels of the two-stage 95% LRS- and MLE-based RCIs for $P_A - P_B$ with Pocock's boundary

P_A	P_B	$n = 5$		$n = 10$		$n = 15$	
		LRS	*MLE*	*LRS*	*MLE*	*LRS*	*MLE*
0.9	0.1	97.9	65.7	96.2	83.6	93.0	91.2
0.8	0.3	90.4	85.2	93.8	86.3	93.7	92.5
0.6	0.4	90.9	88.3	93.8	89.8	94.6	93.1
0.5	0.5	91.8	87.6	93.8	93.8	93.6	93.5

the likelihood ratio test. To compare MLE- and LRS-based RCIs, extensive empirical studies were conducted. In general, with equal increments in information between analyses, LRS intervals perform better than their MLE counterparts for small and moderate-sized samples. For example, Table 11 gives the empirical levels of the 95% LRS- and MLE-based RCIs for the difference Δ between the success probabilities of two treatments with two looks. At each look, n patients are assigned to each treatment. Each entry in the table is based on 10 000 replications. As shown in Table 11, the MLE-based RCIs tend to be too liberal especially when the difference Δ is large. It is important to note that if the looks are not equally spaced in the information scale these approximate RCIs may be inadequate because the critical values may have to be estimated from the data.

For binary data, exact RCIs can be constructed through a network algorithm similar to those of Wei (1988) and Wei *et al.* (1989). Such procedures can readily handle the problems of unequal group sizes and unconventional designs. A full account of this approach will be given in a subsequent paper.

Professor Thomas A. Louis (University of Minnesota, Minneapolis): Jennison and Turnbull do an excellent job in defining repeated confidence intervals (RCIs) and in giving examples of their derivation, use and efficiency for frequentist analysis in a wide variety of clinical trials. However, they give rather less attention to the basic benefits of the RCI approach, and many readers may come away from the article thinking that this approach adds little to the use of repeated significance tests. RCIs produce benefits that are not readily obtainable using tests, including added flexibility in timing and performing interim analyses, a natural method for producing valid confidence intervals once the trial has been terminated and the identification of key parameters or predictions that measure therapeutic effects. Jennison and Turnbull at least mention the first two, and readers acquainted with the difficulties of administering a monitoring committee meeting or constructing confidence intervals once the trial is stopped will be convinced of the superiority of the RCI method. My third point relates to the desirability that statisticians and clinicians establish common scientific ground when designing and monitoring clinical experiments. Use of confidence intervals induces structure and subject area relevance that may be missing if monitoring is guided by hypothesis tests. Mathematical equivalence between the two approaches does not necessarily translate into scientific equivalence.

As with use of hypothesis tests, RCIs still require specification of the nominal level, the error spending function, and monitoring frequency, but also require a more broadly valid parametric or semiparametric model. So, they demand more of the research team and can be less robust. However, the benefits outweigh these drawbacks as long as sufficient time is spent tuning the approach to a specific application. If considerable time is to be spent on such tuning, a more formal Bayesian approach becomes attractive, especially if robustness to prior misspecification is included. Researchers are conducting such robust Bayes trials. The trials and monitoring plans are designed to produce conclusive results for a group of priors (either through use of a mixture prior or by requiring that the most 'intransigent' prior be won over), and I await news of the impact of these trials.

The use of RCIs will produce better documented and organized and credible interim and final analyses. Their advocacy and use may help to close the gap between the practice of clinical trials and the promise of Bayesian theory. I thank the authors for helping to produce these pay-offs.

Dr J. N. S. Matthews (University of Newcastle upon Tyne): The discussion of practical matters in this interesting paper is against the background of a large, possibly multicentre, trial complete with a formal data monitoring committee. The importance of interim analyses in such trials is clear; in this context this paper could be thought of as adapting confidence intervals for interim analyses.