Biometrika (1990), 77, 4, *pp.* 875-92
Printed in Great Britain

Parameter estimation following group sequential hypothesis testing

BY SCOTT S. EMERSON

Epidemiology and Biometry Program, Arizona Cancer Center, Tucson, Arizona 85724, U.S.A.

AND THOMAS R. FLEMING

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.

SUMMARY

Parameter estimation techniques which fail to adjust for the interim analyses of group sequential test designs will introduce bias in much the same way that the repeated use of single sample hypothesis testing causes inflation of the type one statistical error rate. Methods based on the duality of hypothesis testing and interval estimation require definition of an ordering for the outcome space for the test statistic. In this paper, estimation following a group sequential hypothesis test for the mean of a normal distribution with known variance is investigated. A proposed ordering of the sample space based on the maximum likelihood estimate of the mean is found to result in estimates which compare favourably with estimates computed from orderings investigated by Tsiatis, Rosner & Mehta (1984) and Chang & O'Brien (1986) for a variety of group sequential designs. The proposed ordering is then adapted for use when the sizes of groups accrued between analyses is random.

Some key words: Clinical trial; Confidence interval; Estimation; Group sequential; Unequal group sizes.

1. INTRODUCTION

Randomized clinical trials often use interim analyses to address the ethical considerations involved in assessing the efficacy of a medical treatment. Repeated analysis of data as they accrue facilitates the rapid application of experimental findings in all patient populations. It is widely recognized that such analyses must be performed using a formal sequential test to avoid the inflation of the type one statistical error associated with repeated application of fixed sample hypothesis tests. Practicality most often demands a group sequential approach, in which analyses are performed after groups of observations have accrued.

A wide variety of group sequential designs has been proposed that result in appropriately sized tests while providing adequate treatment of the many concerns inherent in a clinical trial, e.g. Pocock (1977, 1982), O'Brien & Fleming (1979), DeMets & Ware (1980, 1982), Lan & DeMets (1983), Whitehead & Stratton (1983), Fleming, Harrington & O'Brien (1984) and Emerson & Fleming (1989). As the focus of this paper is not toward the design of group sequential tests, we provide general notation encompassing most previously proposed designs. We are primarily concerned with the case of a normally

distributed response variable with known variance. This case is applicable, through asymptotic theory, to a wide variety of situations as discussed by Whitehead (1983).

We consider a group sequential design in which we have potential independent observations $Y_{ij} \sim N(\mu, \sigma^2)$, for $j = 1, \ldots, n_i$, $i = 1, \ldots, m$, and σ^2 is known. In this notation, the first subscript numbers the groups accrued between analyses, and the second subscript numbers the observations within the groups. For $k = 1, \ldots, m$, we define the statistics

$$S_k \equiv \sum_{i=1}^{k} \sum_{j=1}^{n_i} Y_{ij}.$$

To specify the group sequential test, we partition the outcome space for S_k into continuation set C_k and stopping sets $S_k^{(0)}$ and $S_k^{(1)}$, for $k = 1, \ldots, m$. Beginning with $k = 1$, if $S_k \in C_k$ we continue our experiment to observe S_{k+1}. The constraint $C_m \equiv \varnothing$, the empty set, guarantees that the study terminates by the mth analysis. We define test statistics $M \equiv \min\{k : S_k \notin C_k\}$, and $S \equiv S_M$. The event $\{M = k\}$ thus corresponds to stopping a study after the kth analysis and rejecting or failing to reject the null hypothesis according to $S \in S_M^{(1)}$ or $S \in {}^{(0)}_{M}$, respectively. For particular choices of continuation and stopping sets, we can write the density for the test statistic (M, S) in the manner of Armitage, McPherson & Rowe (1969) as

$$p(k, s; \mu) = \begin{cases} f(k, s; \mu) & (s \notin C_k), \\ 0 & \text{otherwise}, \end{cases}$$

with $f(k, s; \mu)$ defined recursively by

$$f(1, s; \mu) = \frac{1}{n_1^{\frac{1}{2}}\sigma} \phi\left(\frac{s - n_1\mu}{n_1^{\frac{1}{2}}\sigma}\right),$$

$$f(k, s; \mu) = \int_{C_{k-1}} \frac{1}{n_k^{\frac{1}{2}}\sigma} \phi\left(\frac{s - u - n_k\mu}{n_k^{\frac{1}{2}}\sigma}\right) f(k-1, u; \mu)\, du \quad (k = 2, \ldots, m),$$

where $\phi(x) = (2\pi)^{-\frac{1}{2}} \exp(-x^2/2)$ is standard normal density. The computationally useful relationship

$$p(k, s; \mu) = p(k, s; 0) \exp\left(\frac{s\mu}{\sigma^2} - \frac{\mu^2}{2\sigma^2} \sum_{i=1}^{k} n_i\right) \tag{1}$$

can be easily derived. By integrating this density numerically for specified continuation and stopping sets in a search, one can find a group sequential test design which avoids inflation of the type one error.

A problem analogous to inflation of the type one error arises when using data from a group sequential test to estimate population parameters. In a group sequential design, the study is stopped preferentially when extreme data have been observed. Thus we expect that the usual fixed sample estimators are biased toward the extremes. This potential early stopping also causes the usual fixed sample confidence intervals to have coverage probabilities different from their nominal level. In group sequential testing, the true coverage probability of nominal 90% confidence intervals constructed using the

usual fixed sample methods,

$$\frac{S}{\sum n_i} \pm \frac{1{\cdot}645\sigma}{(\sum n_i)^{\frac{1}{2}}},$$

where the sums are over the range $i = 1, \ldots, M$, can result in coverage probabilities either above or below the nominal level. Numerical computations for the case of a symmetric one-sided test (Emerson & Fleming, 1989) having a maximum of $m = 4$ analyses, equal group sizes, an $\alpha = 0{\cdot}05$ level of significance, and boundary relationships similar to those of Pocock (1977), for example, result in naive 90% confidence intervals having coverage probability ranging from 85% to 93%, depending on the true value of the mean. The behaviour of these intervals is generally worse for the Pocock (1977) boundary relationships than for the O'Brien & Fleming (1979) boundary relationships.

In the present paper, we propose a method of parameter estimation which accounts for the sequential testing, and compare this technique to previously reported methods. We assume that the only information available to us is the sufficient statistic (M, S) from some specified group sequential hypothesis test.

Our method is based on the duality of hypothesis testing and confidence interval estimation. In § 2, we discuss the need for defining an order for the outcome space of the experiment when constructing confidence intervals using data from a group sequential test. We define the proposed ordering and describe two previously reported orderings which will be used for comparison. The use of orderings of the outcome space for point estimation and computation of P-values is the subject of § 3. We compare the behaviour of these three orderings when testing at intervals of equal information in § 4, and in § 5 consider the effect of random number and timing of analyses. The application of these methods to simulated survival data is presented in § 6. We summarize our results in § 7.

2. Exact confidence intervals

A confidence interval is defined as the set of all hypothesized values for a parameter which would not be rejected using an appropriately sized hypothesis test. That is, suppose we have a test statistic T with distribution dependent upon an unknown parameter $\theta \in \Theta$. For a given significance level, α, and a hypothesized parameter value, θ, we determine an acceptance region $\mathscr{A}_\theta^\alpha$, subject to the constraint that $\text{pr}(T \in \mathscr{A}_\theta^\alpha | \theta) = 1 - \alpha$. Our level α test is then defined by:

reject H_0: $\theta = \theta_0$ if and only if $T \notin \mathscr{A}_{\theta_0}^\alpha$.

We invert such a test to find a $(1-\alpha)$ confidence set, $I^\alpha(T)$, by defining $I^\alpha(T) = \{\theta: T \in \mathscr{A}_\theta^\alpha\}$. Under appropriate restrictions on the choice of the acceptance regions, $\{\mathscr{A}_\theta^\alpha: \theta \in \Theta\}$, this confidence set will be an interval. In fixed sample testing, the usual choice is of the form $\mathscr{A}_\theta^\alpha = \{t: \text{pr}(T \geqslant t | \theta) > \alpha\}$. The practicality of such a choice is dependent upon whether, for fixed t, the value of $\text{pr}(T \geqslant t | \theta)$, the probability of observing more extreme results, is nondecreasing as θ gets farther from some null hypothesis Θ_0 in the direction of an alternative Θ_1.

Construction of confidence intervals by test inversion is more complicated in the setting of sequential testing. In this case, the distribution of the test statistic is affected by the testing process itself. Since the sequential test design is chosen based on the null and

alternative hypotheses of interest, we cannot truly examine the results of a similar test of other hypotheses. Had we been interested in other hypotheses, we might have stopped the test earlier or later than we did. Instead we define acceptance regions based on non-'extreme' results for the sufficient statistic, (M, S), and use them to define $(1-\alpha)$ confidence sets, $I^{\alpha}(M, S)$, according to

$$\mathscr{A}_{\mu}^{\alpha} = \{(k, s): \text{pr}\,[(M, S) \geqslant (k, s)|\mu] > \alpha\}, \quad I^{\alpha}(M, S) = \{\mu: (M, S) \in \mathscr{A}_{\mu}^{\alpha}\}. \tag{2}$$

Clearly for these definitions to make sense, we must define an ordering on the outcome space

$$\{(k, s): s \in S_k^{(0)} \cup S_k^{(1)}; k = 1, \ldots, m\}.$$

The best choice for such an ordering is not as clear. The densities for the group sequential statistics lack monotone likelihood ratio, so the theory regarding uniformly most powerful tests and uniformly most accurate confidence bounds (Lehmann, 1959) does not apply. Thus we propose intuitively reasonable methods of ordering the outcome space and investigate the behaviour of such an ordering with respect to various criteria. Every such ordering does not result in true confidence intervals. We return to this point in § 4.

Motivated by the one-parameter families of group sequential designs investigated by Wang & Tsiatis (1987) and Emerson & Fleming (1989), Emerson, in an unpublished dissertation at the University of Washington, investigated a family of orderings, indexed by a parameter q: $(M_1, S_1) < (M_2, S_2)$ if

$$S_1 \Big/ \left(\sum_{i=1}^{M_1} n_i\right)^q < S_2 \Big/ \left(\sum_{i=1}^{M_2} n_i\right)^q.$$

In this paper, we focus on the choice $q = 1$, which is ordering by the sample mean. This was observed to be optimal within the larger family of orderings with respect to a number of criteria.

We compare the proposed ordering to two orderings previously proposed in the literature. In the setting of group sequential tests for a normal mean, Tsiatis, Rosner & Mehta (1984) studied an ordering based primarily on the analysis time at which the study terminates. This ordering was based on the method proposed by Armitage (1957), which was explored by Siegmund (1978) for the case of continuous monitoring and investigated in the setting of a group sequential test of a binomial proportion by Jennison & Turnbull (1983). In this ordering, results corresponding to earlier termination are more extreme than those which terminate later. Ordering of results with common stopping times is done in the natural manner. For a one-sided group sequential test or a two-sided group sequential test in which the continuation sets are intervals, the Tsiatis ordering can be formally defined by $(M_1, S_1) < (M_2, S_2)$ if $M_1 < M_2$ and $S_1 \in S_{M_1}^{(0)}$, or if $M_1 = M_2$ and $S_1 < S_2$, or if $M_1 > M_2$ and $S_2 \in S_{M_2}^{(1)}$. This ordering is not defined for designs with non-interval continuation regions such as the two-sided triangular tests of Whitehead & Stratton (1983) or the two-sided symmetric tests of Emerson & Fleming (1989).

Chang & O'Brien (1986) investigated an ordering based on the likelihood ratio for a group sequential test of a binomial proportion. They ordered the outcome space by defining how extreme a particular value of the test statistic was from a hypothesized parameter value μ_0. For a normal mean, this likelihood ratio ordering would be defined

by the following. For a hypothesized mean, μ_0, (M_1, S_1) was more extreme than (M_2, S_2) if

$$\frac{p(M_1, S_1|\mu = \hat{\mu}_1)}{p(M_1, S_1|\mu_0)} > \frac{p(M_2, S_2|\mu = \hat{\mu}_2)}{p(M_2, S_2|\mu_0)},$$

where $\hat{\mu}_l$ $(l = 1, 2)$ is the maximum likelihood estimate based on (M_l, S_l). Using the form for the density specified by (1), we can specify a total ordering for $\mu = \mu_0$ by $(M_1, S_1) < (M_2, S_2)$ if

$$\left(\sum_{i=1}^{M_1} n_i\right)^{\frac{1}{2}} (\hat{\mu}_1 - \mu_0) < \left(\sum_{i=1}^{M_2} n_i\right)^{\frac{1}{2}} (\hat{\mu}_2 - \mu_0).$$

This ordering is different for each hypothesized value for the parameter. Rosner & Tsiatis (1988) and Chang (1989), in research concurrent to this, have investigated this ordering when applied to the estimation of a normal mean.

3. Point estimation and *P*-values

The concept of ordering the sample space is also useful in some methods of point estimation. One such method of point estimation involves using a median unbiased estimate for the unknown normal mean. Given an observed test statistic $(M, S) = (M^*, S^*)$, one can define an estimate, $\tilde{\mu}$, of the mean by

$$\text{pr}\{(M, S) > (M^*, S^*)|\tilde{\mu}\} = \tfrac{1}{2}.$$

Whitehead (1983) investigated the bias and mean squared error for a median unbiased estimate, $\tilde{\mu}_{Ts}$, based on the Tsiatis ordering of the outcome space. In this paper we investigate the use of a median unbiased estimate, $\tilde{\mu}_{SM}$, based on the sample mean ordering.

Not all point estimates are dependent upon an ordered outcome space. The maximum likelihood estimate of the normal mean, $\hat{\mu}$, is easily found from (1) to be the sample mean. Whitehead (1986) studied the behaviour of a bias adjusted mean estimate, $\breve{\mu}$, defined by

$$E\{(M, S)|\breve{\mu}\} = (M^*, S^*),$$

which is similarly independent of any particular ordering of the outcome space. There also exists a uniform minimum variance unbiased estimator of the sample mean

$$\ddot{\mu} = E\{S_1/n_1|(M^*, S^*)\},$$

where S_1 is the observed partial sum at the time of the first analysis.

Another use of an ordered outcome space is in computing *P*-values for a test result. Under the hypothesis $H_0: \mu = \mu_0$, we might define the *P*-value for an observation (M^*, S^*) as $\text{pr}\{(M, S) > (M^*, S^*)|\mu_0\}$, the probability of obtaining more extreme results than those observed. Clearly, there is not a unique *P*-value for a test result. Instead, we may only

speak of a P-value based on a particular ordering of the outcome space. We do not further address the issues concerning the definition of P-values for a group sequential test.

4. TESTING AT INTERVALS OF EQUAL INFORMATION

We first investigate the relative performance of the proposed estimation techniques in the setting of a group sequential test with groups of equal sizes accrued between analyses, that is $n_i = n$ for $i = 1, \ldots, m$. We assume that m, the maximum number of analyses to be performed, is fixed. We note that the results presented in this section are relatively unaffected by inequalities among the group sizes so long as the exact distribution of the group sequential test statistic, (M, S), is known up to its dependence on the unknown mean, μ. The main use we make of this equal information testing is that the group sizes are known prior to the first analysis.

There are several criteria which we use to compare competing methods for ordering the outcome space. Since analytic results are not always possible, we make these comparisons for specific designs encompassing a spectrum of three important characteristics of group sequential designs: level of significance, maximum number of analyses, and degree of early conservatism, i.e. a spectrum of boundary relationships ranging from those used by O'Brien & Fleming (1979) to those used by Pocock (1977). We base our comparisons in part on representative designs from the family of one-sided symmetric group sequential designs proposed by Emerson & Fleming (1989). In their standardized form, these designs, indexed by maximum number of analyses m, level of significance α, and boundary relationship p, are appropriate for the case where m is fixed in advance, $n_i = 1$ $(i = 1, \ldots, m)$, $Y_{i1} \sim N(\mu, 1)$, and we are testing the hypotheses H_0: $\mu \leqslant -\frac{1}{2}\delta_1$ versus H_1: $\mu \geqslant \frac{1}{2}\delta_1$. The parameter p indexes the degree of early conservatism for the design: $p = 0$ corresponds to O'Brien & Fleming (1979) boundary relationships; $p = \frac{1}{2}$ corresponds to Pocock (1977) boundary relationships.

For those criteria for which numerical results were necessary, the density (1) was numerically integrated for a range of values for the mean using a computer program written in Pascal. Acceptance regions and medians for the orderings described in § 2, as well as the expectation of the sample mean, were tabulated. These tables were then inverted to find confidence bounds and point estimates for any particular value of (M, S), using linear interpolation to find the bounds for those values not explicitly given in the table. The interval of tabulation was small enough to provide coverage probabilities accurate to 0·001.

We must verify that a given ordering of the sample space produces true intervals when applied in (2). In his unpublished dissertation, Emerson proves that the sequential test statistic, (M, S), is stochastically ordered under the sample mean ordering, and also corrects Kim & DeMets (1987) proof of the stochastic ordering of the sequential test statistic under the Tsiatis ordering. Thus, each of these orderings will result in confidence intervals. Several instances were detected numerically, however, in which the confidence set based on the likelihood ratio was not an interval. These departures from true intervals were generally negligible, and we shall continue to refer to the confidence sets under the likelihood ratio ordering as if they were true intervals.

It is also shown in Emerson's dissertation that sample mean based confidence intervals will agree with test results for all practical group sequential test designs in the sense that

a $(1-2\alpha)$ confidence interval will not include hypotheses which are rejected by a one-sided level α group sequential test. This property also holds for the Tsiatis ordering, but not for likelihood ratio based confidence intervals. Table 1 presents 0·9 confidence intervals for selected values of (M, S) from one-sided symmetric group sequential designs with $m=4$, $\alpha=0{\cdot}05$, and $p=0$ and 0·5. Of particular interest are the confidence intervals for the observation (4, 0) which in each case corresponds to observing the border between $S_4^{(0)}$ and $S_4^{(1)}$, the stopping sets at the final analysis. The confidence intervals for both the Tsiatis and sample mean based orderings correspond to the desired value of $(-\frac{1}{2}\delta, \frac{1}{2}\delta)$. The likelihood ratio confidence intervals, on the other hand, are larger than this for the O'Brien & Fleming boundary relationships and smaller than this for the Pocock boundary relationships.

Figure 1 shows the average length of confidence intervals constructed under the methods of § 3 for one-sided symmetric designs having a maximum of $m=6$ analyses, an $\alpha=0{\cdot}05$ level of significance, and having boundary relationships specified by $p=0$, 0·25 and 0·5. The true value of the normal mean is measured by the power of the test to detect that mean. The results for each ordering are expressed as the proportion of the average length of the intervals based on the Tsiatis ordering for each value of the mean. These graphs are representative of the types of general trends in the relative performance of the various methods of estimation. Different designs do exhibit some variation in the ranges for which a specific ordering performs best with respect to average confidence interval length. The sample mean ordering was observed to behave better than the Tsiatis ordering for all designs and all values of the mean, and tends to be better than the likelihood ratio ordering for increasing m and α, and decreasing p. These latter patterns are more evident

Table 1. *Point and interval estimates for selected outcomes of one-sided symmetric group sequential tests* ($m=4$, $\alpha=0{\cdot}05$, $p=0, 0{\cdot}5$)

		Point estimates				90% confidence intervals		
$\hat{\mu}$	M	$\tilde{\mu}_{TS}$	$\tilde{\mu}_{SM}$	$\breve{\mu}$	$\ddot{\mu}$	Tsiatis	LR	Sample mean
				$p=0$: O'Brien & Fleming (1979) boundary relationship				
−3	1	−3·00	−2·88	−2·80	−3·00	(−4·65, −1·36)	(−4·60, −1·33)	(−4·48, −1·34)
	2	−2·47	−2·88	−2·80	−2·12	(−3·86, −0·91)	(−4·05, −1·88)	(−4·48, −1·34)
	3	−0·85	−2·88	−2·80	−0·77	(−2·02, 0·32)	(−3·87, −2·08)	(−4·48, −1·34)
	4	−0·25	−2·88	−2·80	−0·20	(−1·22, 0·70)	(−3·76, −2·19)	(−4·48, −1·34)
−1	2	−1·00	−0·90	−0·84	−0·97	(−2·16, 0·17)	(−2·13, 0·26)	(−2·00, 0·20)
	3	−0·75	−0·90	−0·84	−0·57	(−1·79, 0·33)	(−1·91, 0·01)	(−2·00, 0·20)
	4	−0·25	−0·90	−0·84	−0·15	(−1·22, 0·70)	(−1·80, −0·12)	(−2·00, 0·20)
0	4	0·00	0·00	0·00	0·00	(−0·85, 0·85)	(−0·89, 0·89)	(−0·85, 0·85)
				p=0·5: Pocock (1977) boundary relationship				
−3	1	−3·00	−3·00	−2·97	−3·00	(−4·65, −1·36)	(−4·65, −1·48)	(−4·65, −1·36)
	2	−1·05	−3·00	−2·97	−0·80	(−2·65, −0·64)	(−4·16, −1·89)	(−4·65, −1·36)
	3	−0·35	−3·00	−2·97	−0·18	(−1·54, 0·90)	(−3·95, −2·10)	(−4·65, −1·36)
	4	−0·05	−3·00	−2·97	−0·02	(−1·12, 0·97)	(−3·82, −2·24)	(−4·65, −1·36)
−1	2	−0·75	−0·75	−0·79	−0·45	(−2·02, 0·67)	(−2·09, 0·25)	(−2·00, 0·64)
	3	−0·35	−0·75	−0·79	−0·12	(−1·50, 0·90)	(−1·90, 0·02)	(−2·00, 0·64)
	4	−0·05	−0·75	−0·79	−0·01	(−1·12, 0·98)	(−1·78, −0·12)	(−2·00, 0·64)
0	4	0·00	0·00	0·00	0·00	(−1·01, 1·01)	(−0·92, 0·92)	(−1·01, 1·01)

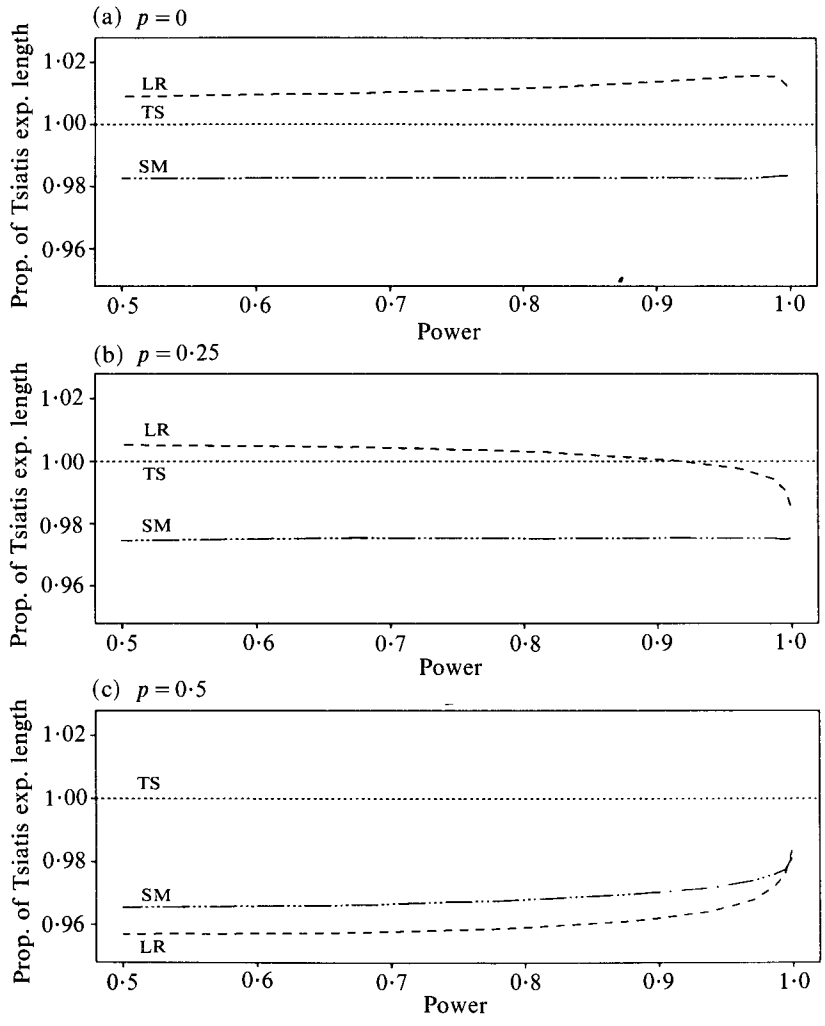

Fig. 1. Comparison of average length of 0·9 confidence intervals for one-sided symmetric tests with selected boundary relationships. Average length of confidence intervals is expressed as a proportion of the expected length of Tsiatis intervals and is displayed versus the power of the test to detect the true mean for group sequential designs having parameters $m=6$, $\alpha=0·05$, and $p=0$, 0·25 and 0·5. TS, Tsiatis; LR, likelihood ratio based; SM, sample mean based.

under the intermediate hypothesis than more extreme hypotheses. Nonsymmetric group sequential designs also showed these trends.

Figure 2 compares the bias and mean squared error for $\tilde{\mu}_{SM}$, the median unbiased estimate based on the sample mean ordering, to $\tilde{\mu}_{TS}$, the median unbiased estimate based on the Tsiatis ordering, the maximum likelihood estimate $\hat{\mu}$, the bias adjusted mean $\breve{\mu}$, and the uniform minimum variance unbiased estimate $\ddot{\mu}$. Similar results are observed for all other designs investigated. The bias adjusted mean has the least absolute bias of the biased estimators and lowest mean squared error of all the estimators almost uniformly over the range of alternatives for which the test has power between 0·01 and 0·99. This then raises an additional question regarding agreement between the various confidence intervals and that point estimator. From Table 1 the various confidence intervals differ

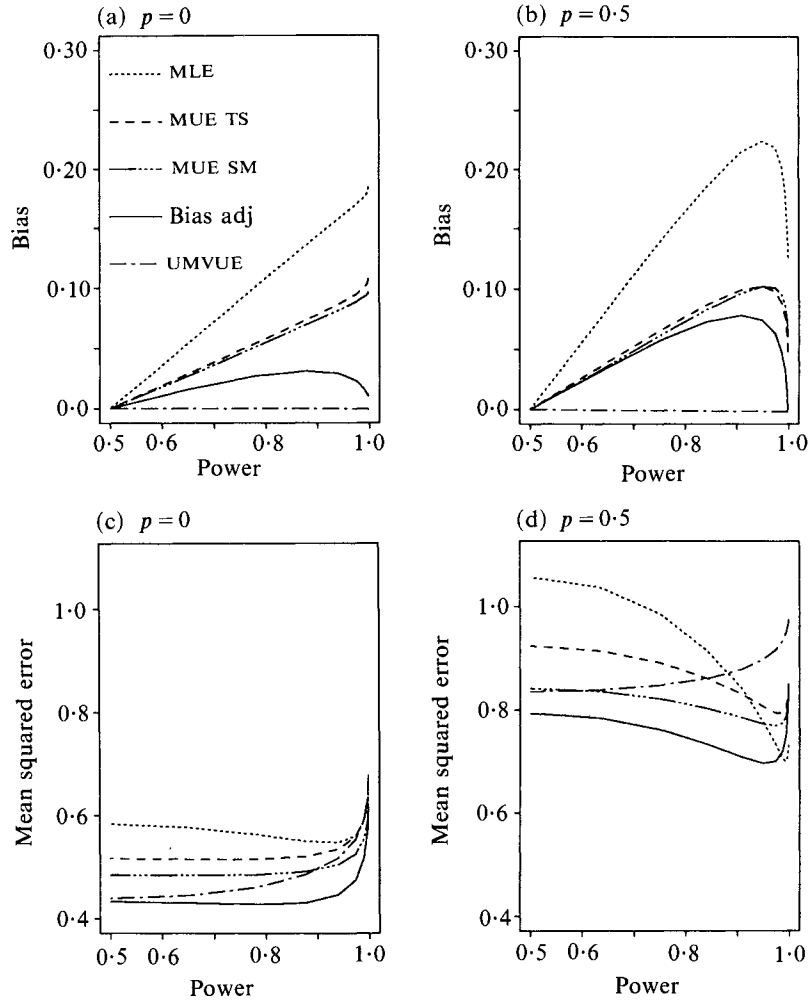

Fig. 2. Bias and mean squared error of point estimators versus power of test to detect the true mean for standardized one-sided symmetric group sequential designs with parameters $m=4$, $\alpha=0{\cdot}05$, and $p=0$ and $0{\cdot}5$. MLE, maximum likelihood estimate; MUE TS, Tsiatis based median unbiased estimator; MUE SM, sample mean based median unbiased estimator; Bias adj, bias adjusted estimator; UMVUE, uniform minimum variance unbiased estimator.

with respect to inclusion of the proposed point estimates. The likelihood ratio and sample mean based intervals behave generally alike with respect to this criterion: they will include the maximum likelihood estimate, the sample mean based median unbiased estimate, and the bias adjusted estimator, but occasionally will not include the uniform minimum variance unbiased estimate. The Tsiatis based ordering will include the Tsiatis based median unbiased estimate and the uniform minimum variance unbiased estimate, though for higher values of M, it will not always contain the maximum likelihood estimate or the bias adjusted estimator. The exact behaviour of the intervals with respect to this criterion will depend upon the test design used. The frequency with which this problem arose was observed to be less than 1% for all designs investigated, with better behaviour

for test designs corresponding to O'Brien & Fleming boundary relationships compared to those with Pocock boundary relationships.

5. RANDOM TIMING OF ANALYSES

The results presented in § 4 are applicable when the groups accrued between successive analyses are of equal size. In many clinical trials, this is not the case. Data monitoring committees often meet at regular intervals of calendar time, and analyses of the data are performed in preparation for these meetings. Random patterns of data accrual and delayed measurement of response result in a nonuniform rate of information accrual. Such a setting entails three important departures from the assumptions of the previous section: (i) the maximum number of analyses to be performed may be a random variable, (ii) the sizes of the groups accrued between analyses may be unequal and may not be known prior to the time of each analysis, and (iii) the maximal sample size accrued may vary from that which was planned. In this section we discuss the impact each of these departures has on the proposed method of computing confidence intervals and investigate an adaptation of this method to the setting of random timing of analyses.

Several solutions have been proposed to the problems posed by random group sizes to the specification of group sequential test designs. While exact methods have been proposed by Lan & DeMets (1983) and Fleming et al. (1984), among others, it has been found that the use of tests appropriate for testing at equal information will often be sufficiently correct when group sizes are unequal (Pocock, 1977; DeMets & Gail, 1985; Emerson & Fleming, 1989). Unfortunately, this naive use of equal information methods is not so robust when it comes to estimation. Confidence intervals derived using the density for equal information testing can have coverage probabilities markedly above or below the desired level when used in the setting of unequal group sizes. For some group sequential designs, the true coverage probability for nominal 0·9 confidence intervals was observed to range between 0·75 to 0·98 as the true value of the normal mean varied. Similarly, the various point estimators exhibited as much as a ten-fold increase in bias when mis-applied in this fashion. These fluctuations were generally worse for the Pocock type boundary relationships than the O'Brien & Fleming type boundary relationships.

We must therefore use methods which account for the true group sizes accrued between analyses when computing estimates from group sequential data. However, we need to know the entire density for the group sequential test statistic to compute some of the estimators. Both the sample mean ordering and the likelihood ratio ordering are dependent upon the prior knowledge of all group sizes. An analogous problem exists with the computation of the bias adjusted mean, but not the uniform minimum variance unbiased estimate or the estimators based on the Tsiatis ordering. In the case of the latter, more 'extreme' results are only associated with analyses prior to the observed stopping at the Mth analysis, and thus all group sizes affecting the computation will have been observed.

Our approach to this problem relies on a feature common to many of the techniques for specifying group sequential test designs when the timing of analyses is random. In many of these methods, e.g. Lan & DeMets (1983) and Emerson & Fleming (1989), the maximum sample size to be accrued, $n_0 = n_1 + \ldots + n_m$, is specified in advance, though the number of groups, m, and the individual group sizes, n_i $(i = 1, \ldots, m)$ need not be. Our rule then is as follows. Whenever we terminate a study early, we assume that $m = M+1$ and that $n_m = n_0 - (n_1 + \ldots + n_M)$. That is, we assume that only a single additional analysis would have been performed, and that that analysis would have occurred when

the maximum sample size, n_0, had been accrued. We base our estimates on a density for (M, S) using these group sizes. We note that this adaptation is in some respects a worst case assumption. Better results are to be expected if we were to use a more realistic estimate of m, perhaps based on the original design, and assume that all remaining group sizes were equal.

To investigate the robustness of this adaptation, we compare the approximate estimators based on the sample mean ordering to the exact estimates based on the Tsiatis ordering. Our goal is to demonstrate that this approximation behaves well under a spectrum of accrual patterns and thus performs adequately in the setting of truly random group sizes when the actual accrual pattern might be chosen randomly from the patterns investigated.

We consider a family of accrual patterns, parameterized by r, in which the sizes of the groups accrued between analyses obeys the relation

$$\sum_{i=1}^{k} n_i = \left(\frac{k}{m}\right)^r n_0.$$

In this family, accrual patterns with $r<1$ correspond to initially faster accrual which slows as the study progresses. When $r>1$, the earlier group sizes are smaller compared to those at later analysis times. This latter case is of most interest since it tends to arise when the measurement of response is delayed, such as occurs in survival analysis. We note that $r=1$ is the case of testing at equal information.

Our comparisons are again based on numerical results for exact, one-sided symmetric group sequential designs for selected values of m, α and p (Emerson & Fleming, 1989). We note that, while the group sequential design used will undoubtedly affect the relative performance of the approximation, the method of applying that design, be it according to Lan & DeMets (1983), Fleming et al. (1984), or Emerson & Fleming (1989), is not important. In these comparisons, we compute estimates based on the approximation given above, but compute coverage probabilities, expected length, bias and mean squared error using the exact density under the true, but hypothetically unknown, accrual pattern.

Figure 3 displays the actual coverage probability of confidence intervals based on the sample mean ordering when using the approximation for random group sizes. The behaviour of these intervals is generally quite good, especially for designs with O'Brien & Fleming boundary relationships ($p=0$), when $r>1$, and for alternatives for which the test has power between 0·01 and 0·99. These intervals were observed to continue to perform better than the Tsiatis based intervals with respect to expected length for all values of the mean.

Results for the point estimators based on the approximate distribution are presented in Table 2. From this table it can be seen that the bias adjusted mean continues to have low bias and low mean squared error. It is interesting to note that the estimator based on the approximate distribution often outperforms the analogous estimator using the true distribution.

The approximations used in this setting were also found to be relatively robust to departures from the setting of fixed n_0. Numerical results for tests which have maximum sample sizes 10% higher or lower than the value used in the approximation were found to have the same good performance. Coverage probabilities of 0·9 confidence intervals were within ±0·005 of the nominal level for the range of alternatives for which the power of the test was 0·01 to 0·99. Bias and mean squared error of point estimators continued to show improvement over the exact methods of the Tsiatis based median unbiased

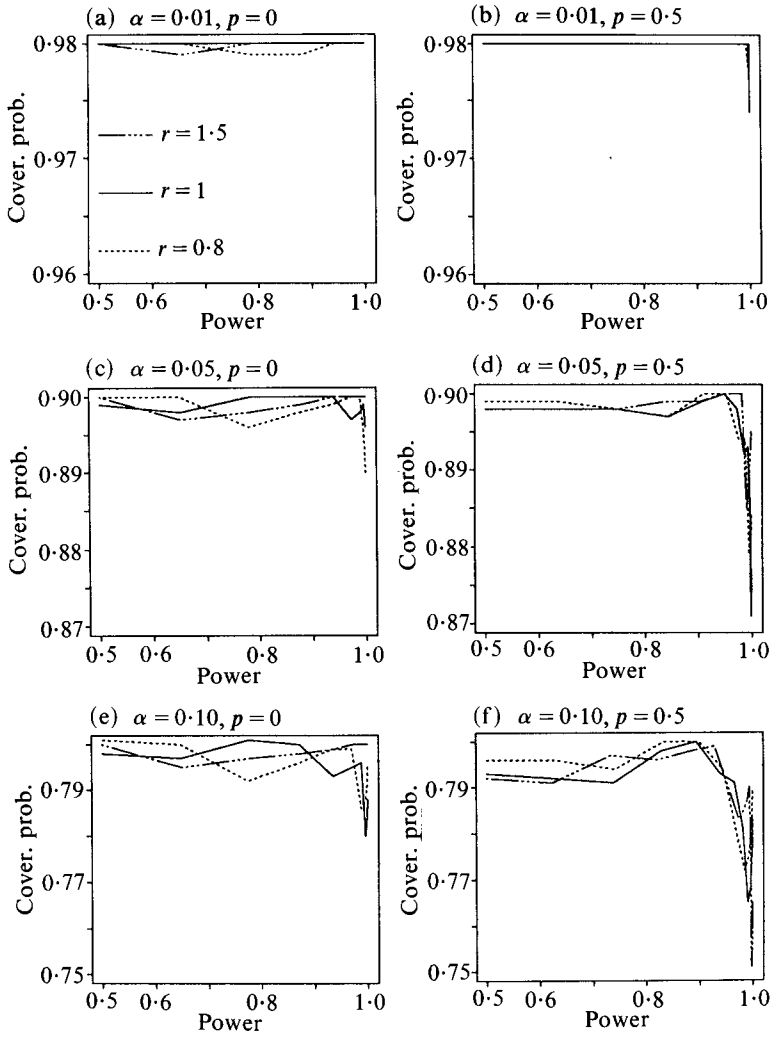

Fig. 3. Coverage probability of $(1-2\alpha)$ confidence intervals from derived tests under varying accrual patterns. Test designs used are one-sided symmetric group sequential designs with parameters $m = 4$, $\alpha = 0{\cdot}01$, $0{\cdot}05$ and $0{\cdot}10$, and $p = 0$ and $0{\cdot}5$, for data accrual rates corresponding to $r = 0{\cdot}8$, $1{\cdot}0$ and $1{\cdot}5$.

estimate, and the mean squared error of the approximation to the bias adjusted mean was less than that of the uniform minimum variance unbiased estimate.

6. Application to survival analysis

We apply the above methods to a proportional hazards model (Cox, 1972) in which the survival time of the ith patient has survival distribution $\{S_0(t)\}^{\exp(Z_i\beta)}$, where Z_i is an indicator variable measuring whether the ith patient received the new treatment, β is the parameter measuring treatment effect which is to be estimated, and $S_0(t)$ is some unspecified baseline survivor function. For the kth analysis, let $T_i^{(k)}$ be the observation time for the ith patient, $d_i^{(k)}$ indicate an observed death, and $r_{0i}^{(k)}$ and $r_{1i}^{(k)}$ be the number of patients in the control and treatment groups, respectively, who are observed to survive

Table 2. *Bias (mean squared error) of point estimates under derived tests and full design for selected patterns of accrual; one-sided symmetric designs, $m = 4$*

True mean	Power	Derived tests $\tilde{\mu}_{SM}$	Derived tests $\breve{\mu}$	Full design $\tilde{\mu}_{SM}$	Full design $\breve{\mu}$
		$p = 0, r = 0{\cdot}8$			
0·00	0·500	0·000 (0·47)	0·000 (0·43)	0·000 (0·49)	0·000 (0·45)
0·40	0·780	0·044 (0·47)	0·023 (0·42)	0·052 (0·49)	0·030 (0·44)
1·20	0·989	0·071 (0·51)	0·009 (0·48)	0·089 (0·53)	0·034 (0·51)
1·60	0·999	0·064 (0·55)	−0·006 (0·54)	0·085 (0·56)	0·028 (0·56)
2·40	1·000	0·033 (0·64)	−0·037 (0·65)	0·055 (0·63)	0·005 (0·63)
		$p = 0, r = 1{\cdot}5$			
0·00	0·500	0·000 (0·46)	0·000 (0·42)	0·000 (0·47)	0·000 (0·43)
0·40	0·783	0·050 (0·46)	0·026 (0·41)	0·052 (0·47)	0·030 (0·43)
1·20	0·990	0·081 (0·49)	0·019 (0·46)	0·090 (0·51)	0·034 (0·48)
1·60	0·999	0·071 (0·54)	0·000 (0·51)	0·085 (0·55)	0·022 (0·52)
2·40	1·000	0·042 (0·67)	−0·048 (0·65)	0·057 (0·67)	−0·019 (0·64)
		$p = 0{\cdot}5, r = 0{\cdot}8$			
0·00	0·500	0·000 (0·65)	0·000 (0·63)	0·000 (0·68)	0·000 (0·66)
0·40	0·752	0·036 (0·64)	0·037 (0·60)	0·045 (0·67)	0·047 (0·63)
1·20	0·976	0·042 (0·63)	0·012 (0·57)	0·063 (0·63)	0·037 (0·58)
1·60	0·995	0·024 (0·66)	−0·021 (0·64)	0·043 (0·65)	0·006 (0·64)
2·40	1·000	−0·002 (0·74)	−0·041 (0·79)	0·009 (0·72)	−0·022 (0·76)
		$p = 0{\cdot}5, r = 1{\cdot}5$			
0·00	0·500	0·000 (1·33)	0·000 (1·13)	0·000 (1·36)	0·000 (1·20)
0·40	0·748	0·093 (1·31)	0·058 (1·11)	0·098 (1·34)	0·068 (1·18)
1·20	0·969	0·183 (1·24)	0·075 (1·07)	0·198 (1·27)	0·106 (1·14)
1·60	0·990	0·185 (1·26)	0·049 (1·13)	0·203 (1·29)	0·090 (1·19)
2·40	0·999	0·153 (1·37)	−0·008 (1·35)	0·175 (1·39)	0·047 (1·36)

at least as long as $T_i^{(k)}$. From the partial likelihood we compute the score and information as

$$U^{(k)}(\beta) = \sum_{i=1}^{N} d_i^{(k)} \left(Z_i - \frac{r_{1i}^{(k)} e^{\beta}}{r_{0i}^{(k)} + r_{1i}^{(k)} e^{\beta}} \right),$$

$$I^{(k)}(\beta) = \sum_{i=1}^{N} d_i^{(k)} \left\{ \frac{r_{1i}^{(k)} e^{\beta}}{r_{0i}^{(k)} + r_{1i}^{(k)} e^{\beta}} \left(1 - \frac{r_{1i}^{(k)} e^{\beta}}{r_{0i}^{(k)} + r_{1i}^{(k)} e^{\beta}} \right) \right\}.$$

Asymptotic results (Andersen & Gill, 1982) show that as the number of observed deaths becomes large, in distribution,

$$U^{(k)}(\beta)\{I^{(k)}(\beta)\}^{-\frac{1}{2}} \to N(0,1), \quad (\hat{\beta}^{(k)} - \beta)\{I^{(k)}(\beta)\}^{\frac{1}{2}} \to N(0,1),$$

where $\hat{\beta}^{(k)}$ satisfies $U^{(k)}(\hat{\beta}^{(k)}) = 0$. From a first order Taylor expansion of $U^{(k)}(0)$ about some small β we thus have $U^{(k)}(0) \sim N\{\beta I^{(k)}(\beta), I^{(k)}(\beta)\}$ for large numbers of deaths. The validity of group sequential analysis using the partial likelihood based score has been verified by Tsiatis (1982), Sellke & Siegmund (1983) and Slud (1984), who addressed the asymptotic independent increment structure of these statistics. Thus we may base a group sequential test on the logrank statistic, $S_k = U^{(k)}(0)$, and use the above approximate distribution under small alternatives to compute confidence intervals and point estimates. Similarly, a first order Taylor expansion of $U^{(k)}(0)$ about $\hat{\beta}^{(k)}$ suggests that we might alternatively base our inference on the Wald statistic $S_k = \hat{\beta}^{(k)} I^{(k)}(\hat{\beta}^{(k)})$.

For either test statistic, we have that S_k is approximately $N\{\beta I^{(k)}(\beta), I^{(k)}(\beta)\}$. For large samples we expect the ratio $r_{0i}^{(k)}/r_{1i}^{(k)}$ to be relatively constant, thus $I^{(k)}(\beta) \doteq \Sigma\, d_i^{(k)}\, \sigma^2$, where the sum is over $i = 1, \ldots, N$. We will define our group sizes, n_i $(i = 1, \ldots, m)$ as the number of additional deaths to be observed between successive analyses. Thus we have for $k = 1, \ldots, m$

$$\sum_{i=1}^{k} n_i = \sum_{i=1}^{N} d_i^{(k)}.$$

If patients are randomized equally to the treatment and control groups, we expect $r_{0i}^{(k)} \doteq e^{\beta} r_{1i}^{(k)}$ and therefore use $\sigma^2 = 0{\cdot}25$ to design a group sequential trial of the hypotheses $H_0\colon \beta = 0$, no treatment effect, versus $H_1\colon \beta \leqslant \beta_1 = 0{\cdot}6931$, a placebo-treatment hazard ratio of 2·0.

Though the analysis of the data can allow more flexible determination of the number and timing of analyses, for the purpose of test design, we need to estimate the number of analyses to be performed, m, as well as the relative sizes of the groups accrued between analyses. We note that, unless there were sufficient prior knowledge to suggest otherwise, test design based on testing after intervals of equal information would seem reasonable. Emerson & Fleming (1989) found that tests designed in this manner would have operating characteristics that are relatively unaffected by slight changes in the actual timing of analyses. Let $n = (n_1 + \ldots + n_m)/m$ be the average group size. Then if we perform the transformation

$$S_k^* = -\frac{S_k}{n^{\frac{1}{2}}\sigma} - \frac{\delta_1}{2n} \sum_{i=1}^{k} n_i$$

where $\delta_1 = -\frac{1}{2} n^{\frac{1}{2}} \beta_1 / 2$, we have that the S_k^* have the correct distribution under the null and alternative hypotheses to be used in the standardized form of the one-sided symmetric tests of Emerson & Fleming (1989) when centred about zero.

We plan group sequential designs having a maximum of $m = 4$ analyses, a level of significance $\alpha = 0{\cdot}05$, and boundary relationships corresponding to $p = 0$, the O'Brien & Fleming boundary relationships, or $p = 0{\cdot}5$, the Pocock boundary relationships. We use values of the standardized alternative appropriate for the specific design, $\delta_1 = 2m^p c_{m,p}^{(\alpha)}/m$, and compute the sample size based on the above relation between δ_1 and the untransformed alternative β_1, $n = 4\delta_1^2/\beta_1^2$, thereby fixing the maximum number of deaths to be observed in the clinical trial. From this, we determine that a maximum of $mn = 97$ deaths are required when $p = 0$ and a maximum of $mn = 135$ deaths are required when $p = 0{\cdot}5$. Continuation and stopping sets for the S_k^*'s chosen according to the methods of § 5 of Emerson & Fleming (1989) yield a group sequential test symmetric in the treatment of H_0 and H_1.

To determine the numbers of patients accrued to the study, we assume uniform accrual over a three year period and one year of additional follow-up. The sample size is then computed using exponentially distributed survival times with mean survival of 1 year in the control group and 2 years in the treatment group. The sample size chosen is that value which would be expected to yield the specified maximum number of deaths, 97 or 135, at the end of the fourth year. For an accrual period of a years, and a total follow-up period, including accrual, of f years, the probability of observing a death $\mathscr{D}$ during the study is given by

$$\text{pr}\,(\mathscr{D} \mid \lambda) = \int_0^a \int_0^{f-v} (a\lambda)^{-1} e^{-u/\lambda}\, du\, dv,$$

where λ is the mean survival time. We find the maximum sample size N such that the specified maximum number of deaths, mn, satisfies

$$\tfrac{1}{2}N\{\text{pr}(\mathscr{D}|\lambda_0)+\text{pr}(\mathscr{D}|c\lambda_0)\},$$

where c is the hazard ratio under the design alternative and λ_0 is the baseline hazard function. Thus for the simulated clinical trial, the required number of patients is 62 per arm for $p=0$ and 86 per arm for $p=0{\cdot}5$. Under this model, the information accrual pattern with $r=1{\cdot}5$ would correspond to analyses at intervals of approximately 1 year.

For particular values of β and p, 5000 clinical trials were simulated according to exponential survival times. For each such simulation, group sequential tests based on the score statistic, with $S_k=U^{(k)}(0)$ and using $I^{(k)}(0)$ to estimate σ^2, and the Wald statistic, with $S_k=\hat{\beta}^{(k)}I^{(k)}(\hat{\beta}^{(k)})$ and using $I^{(k)}(\hat{\beta}^{(k)})$ to estimate σ^2, were performed. For reasons of simulation efficiency, we examine only a single data accrual pattern, though as noted in § 5, the performance of our approximation for random group sizes is relatively independent of the exact accrual rate over a wide range of accrual patterns. Thus we approximate the case of random group sizes since we do not assume knowledge of the actual group sizes for analyses which would have occurred after termination of the study, but we present results which are conditioned on a specific accrual pattern. In the simulations, analyses were performed after observing 12, 34, 63 and 97 deaths for $p=0$, and after observing 17, 48, 88 and 135 deaths for $p=0{\cdot}5$.

Using the results of those tests, and the methods of § 5, various point and interval estimates were computed based on S_M^*. An estimate $\tilde{\delta}$ based on S_M^* corresponds to an estimate of β in the untransformed problem according to

$$\tilde{\beta}=-\frac{(\tilde{\delta}+\frac{1}{2}\delta_1)}{(n^{\frac{1}{2}}\sigma)}.$$

We investigate the behaviour of the proposed estimation techniques for three alternatives: (i) $\beta=0$ corresponding to the null hypothesis of no treatment effect, (ii) $\beta=-0{\cdot}5878$ corresponding to an intermediate hypothesis of a control to treatment hazard ratio of 1·8, and (iii) $\beta=-0{\cdot}6931$ corresponding to the design alternative of a control to treatment hazard ratio of 2·0. The empirical size and power of the tests under the null and alternative hypotheses were found to be extremely close to the desired levels of 0·05 and 0·95, respectively: empirical size ranged from 0·046 to 0·049; empirical power ranged from 0·934 to 0·948.

Table 3 presents the coverage probabilities of the confidence intervals derived using the approximate sample mean ordering, the exact Tsiatis ordering and the naive, fixed sample techniques. From this table it can be seen that the empirical coverage probabilities for the approximate sample mean ordering are indistinguishable from those of the exact Tsiatis method. In Table 4 we present the bias and mean squared error of several point estimators. Evident is the bias of the fixed sample estimate, which is also the maximum likelihood estimate $\hat{\beta}$. For the Wald based test and estimates with $p=0{\cdot}5$ and $\beta=-0{\cdot}6931$, the observed expected value of $\hat{\beta}=-0{\cdot}825$ corresponds to a hazard ratio of 2·28, rather than the true value of 2·0. It can be seen that the bias adjusted mean consistently outperforms the other three biased estimates both in bias and standard error. For this same case, the bias adjusted mean has observed expected value of $\check{\beta}=-0{\cdot}715$, which corresponds to a hazard ratio of 2·04. The bias adjusted mean is also seen to have less variance than the uniform minimum variance unbiased estimate, which translates into a lower mean squared error for the bias adjusted mean. It is interesting to note that there

Table 3. *Observed coverage probability of 90% confidence intervals under specified alternatives; one-sided symmetric designs, $m=4$, $\alpha=0{\cdot}05$, 5000 simulations*

p		β			β		
		0	$-0{\cdot}5878$	$-0{\cdot}6931$	0	$-0{\cdot}5878$	$-0{\cdot}6931$
		Score based intervals			Wald based intervals		
0	Naive	0·8888	0·8514	0·8760	0·8946	0·8630	0·8846
	Tsiatis	0·8990	0·8916	0·8934	0·9044	0·9004	0·9000
	SM	0·8994	0·8922	0·8930	0·9050	0·9004	0·8990
0·5	Naive	0·8772	0·8286	0·8734	0·8852	0·8316	0·8758
	Tsiatis	0·8958	0·9058	0·8992	0·9060	0·9094	0·9052
	SM	0·8990	0·9078	0·9080	0·9080	0·9124	0·9124

Table 4. *Average (standard deviation) of point estimates observed under specified alternatives; one-sided symmetric designs, $m=4$, $\alpha=0{\cdot}05$, 5000 simulations*

p		$\beta=0$	$\beta=-0{\cdot}5878$	$\beta=-0{\cdot}6931$	$\beta=0$	$\beta=-0{\cdot}5878$	$\beta=-0{\cdot}6931$
		Score based intervals			Wald based intervals		
0	$\hat{\beta}$	0·064 (0·30)	−0·664 (0·31)	−0·780 (0·30)	0·065 (0·29)	−0·654 (0·31)	−0·772 (0·31)
	$\tilde{\beta}_{TS}$	0·034 (0·30)	−0·638 (0·30)	−0·749 (0·30)	0·035 (0·29)	−0·629 (0·30)	−0·742 (0·31)
	$\tilde{\beta}_{SM}$	0·028 (0·29)	−0·634 (0·29)	−0·743 (0·29)	0·029 (0·28)	−0·625 (0·29)	−0·734 (0·30)
	$\dot{\beta}$	0·009 (0·27)	−0·619 (0·28)	−0·723 (0·28)	0·009 (0·27)	−0·610 (0·27)	−0·714 (0·28)
	$\ddot{\beta}$	0·000 (0·30)	−0·608 (0·30)	−0·714 (0·31)	0·001 (0·30)	−0·599 (0·30)	−0·706 (0·31)
0·5	$\hat{\beta}$	0·147 (0·40)	−0·681 (0·42)	−0·807 (0·40)	0·152 (0·40)	−0·689 (0·45)	−0·825 (0·44)
	$\tilde{\beta}_{TS}$	0·088 (0·41)	−0·630 (0·41)	−0·742 (0·40)	0·092 (0·41)	−0·637 (0·44)	−0·760 (0·45)
	$\tilde{\beta}_{SM}$	0·073 (0·39)	−0·623 (0·38)	−0·734 (0·37)	0·076 (0·39)	−0·627 (0·41)	−0·746 (0·41)
	$\dot{\beta}$	0·039 (0·36)	−0·604 (0·35)	−0·704 (0·34)	0·042 (0·36)	−0·608 (0·37)	−0·715 (0·37)
	$\ddot{\beta}$	0·015 (0·43)	−0·571 (0·40)	−0·668 (0·41)	0·019 (0·43)	−0·576 (0·43)	−0·680 (0·45)

is a slight bias in the estimate derived from the uniform minimum variance unbiased estimate due to relatively small sample sizes. In the case of the score based estimates with $p=0{\cdot}5$ and $\beta=-0{\cdot}6931$, the bias adjusted estimate has less absolute bias than the uniform minimum variance unbiased estimate.

7. CONCLUSIONS

We have found that a sample mean based ordering of the outcome space for a group sequential test produces exact confidence intervals that perform well in comparison to the previously proposed orderings. The sample mean based ordering gave uniformly average shorter confidence intervals than the ordering investigated by Tsiatis et al. (1984). The former method is also more desirable from the aspect of inclusion of point estimates and application to non-interval continuation sets, such as occur in some previously proposed two-sided test designs.

The sample mean based ordering was also found to perform well in comparison to that based on the likelihood ratio, especially for higher numbers of analyses, higher levels of significance and the designs tending toward more conservative testing at earlier analyses. In addition, the sample mean ordering was found to produce true intervals and to produce intervals consistent with test results. The likelihood ratio ordering was found

to violate both of these desired properties, though such departures were slight in the cases examined in this research.

A point estimate based on the sample mean ordering performed generally better than the analogous estimate based on the Tsiatis ordering. Neither of these estimates performed as well as the bias adjusted estimator investigated by Whitehead (1986), however.

In extending these results to the case of random group sizes, we find that although the group sequential tests themselves are fairly robust to varying accrual rates, the estimation procedures are affected more severely. An adaptation of these methods based on estimating the unobserved group sizes was found to perform quite well, however. The approximate confidence intervals based on the sample mean ordering as well as the approximation to Whitehead's (1986) bias adjusted mean appear to be useful methods in estimation of parameters using data collected in a group sequential test.

Acknowledgement

This work was supported in part by National Institutes of Health, National Cancer Institute grants.

References

ANDERSEN, P. K. & GILL, R. D. (1982). Cox's regression model for counting processes: A large sample study. *Ann. Statist.* **10**, 1100-20.

ARMITAGE, P. (1957). Restricted sequential procedures. *Biometrika* **44**, 9-26.

ARMITAGE, P., MCPHERSON, C. K. & ROWE, B. C. (1969). Repeated significance tests on accumulating data. *J. R. Statist. Soc.* A **132**, 235-44.

CHANG, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45**, 247-54.

CHANG, M. N. & O'BRIEN, P. C. (1986). Confidence intervals following group sequential tests. *Controlled Clin. Trials* **7**, 18-26.

COX, D. R. (1972). Regression models and life-tables (with discussion). *J. R. Statist. Soc.* B **34**, 187-220.

DEMETS, D. L. & GAIL, M. H. (1985). Use of logrank tests and group sequential methods at fixed calendar times. *Biometrics* **41**, 1039-44.

DEMETS, D. L. & WARE, J. H. (1980). Group sequential methods in clinical trials with a one-sided hypothesis. *Biometrika* **67**, 651-60.

DEMETS, D. L. & WARE, J. H. (1982). Asymmetric group sequential boundaries for monitoring clinical trials. *Biometrika* **69**, 661-3.

EMERSON, S. S. & FLEMING, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45**, 905-23.

FLEMING, T. R., HARRINGTON, D. P. & O'BRIEN, P. C. (1984). Designs for group sequential tests. *Controlled Clin. Trials* **5**, 348-61.

JENNISON, C. & TURNBULL, B. W. (1983). Confidence intervals for a binomial parameter following a multistage test with application to MIL-STD 105D and medical trials. *Technometrics* **25**, 49-58.

KIM, K. & DEMETS, D. L. (1987). Estimation following group sequential tests in clinical trials. *Biometrics* **43**, 857-64.

LAN, K. K. G. & DEMETS, D. L. (1983). Discrete sequential boundaries for clinical trials. *Biometrika* **70**, 659-63.

LEHMANN, E. L. (1959). *Testing Statistical Hypotheses.* New York: Wiley.

O'BRIEN, P. C. & FLEMING, T. R. (1979). A multiple testing procedure for clinical trials. *Biometrics* **35**, 549-56.

POCOCK, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64**, 191-9.

POCOCK, S. J. (1982). Interim analyses for randomized clinical trials: The group sequential approach. *Biometrics* **38**, 153-62.

ROSNER, G. L. & TSIATIS, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75**, 723-9.

SELLKE, K. & SIEGMUND, D. (1983). Sequential analysis of the proportional hazards model. *Biometrika* **70**, 315-26.

SIEGMUND, D. (1978). Estimation following sequential tests. *Biometrika* **65**, 341–9.
SLUD, E. V. (1984). Sequential linear rank tests for two-sample censored survival data. *Ann. Statist.* **12**, 551–71.
TSIATIS, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *J. Am. Statist. Assoc.* **77**, 855–61.
TSIATIS, A. A., ROSNER, G. L. & MEHTA, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40**, 797–803.
WANG, S. K. & TSIATIS, A. A. (1987). Approximately optimal one-parameter boundaries for group sequential trials. *Biometrics* **43**, 193–9.
WHITEHEAD, J. (1983). *The Design and Analysis of Sequential Clinical Trials.* Chichester: Ellis Horwood.
WHITEHEAD, J. (1986). On the bias of maximum likelihood estimation following a sequential test. *Biometrika* **73**, 573–81.
WHITEHEAD, J. & STRATTON, I. (1983). Group sequential clinical trials with triangular continuation regions. *Biometrics* **39**, 227–36.

[*Received September* 1989. *Revised April* 1990]