Biometrics 61, 546–551
June 2005

DOI: 10.1111/j.1541-0420.2005.040342.x

A Note on *P*-Values under Group Sequential Testing and Nonproportional Hazards

Daniel L. Gillen

Department of Statistics, University of California, Irvine, California 92697, U.S.A
email: dgillen@uci.edu

and

Scott S. Emerson

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

Summary. Group sequential designs are often used for periodically assessing treatment efficacy during the course of a clinical trial. Following a group sequential test, *P*-values computed under the assumption that the data were gathered according to a fixed sample design are no longer uniformly distributed under the null hypothesis of no treatment effect. Various sample space orderings have been proposed for computing proper *P*-values following a group sequential test. Although many of the proposed orderings have been compared in the setting of time-invariant treatment effects, little attention has been given to their performance when the effect of treatment within an individual varies over time. Our interest here is to compare two of the most commonly used methods for computing proper *P*-values following a group sequential test, based upon the analysis time (AT) and *Z*-statistic orderings, with respect to resulting power functions when treatment effects on survival are delayed. Power under the AT ordering is shown to be heavily influenced by the presence of a delayed treatment effect, while power functions corresponding to the *Z*-statistic ordering remain robust under time-varying treatment effects.

Key words: Censored data; Clinical trials; Nonproportional hazards; Ordering; Sequential tests.

1. Introduction

Time to event is a common outcome in many clinical trials. In such settings, the log-rank score statistic (Mantel, 1966) is typically used to compare the survival experience of two randomly sampled groups. In addition, group sequential testing has become commonplace in clinical trials where the need to achieve a high standard of patient ethics is of utmost importance. Given the independent increments structure of the log-rank statistic (Tsiatis, 1982), group sequential stopping boundaries maintaining a prespecified type I error rate can easily be computed via the sequential density derived by Armitage, McPherson, and Rowe (1969). Such a stopping rule is often the basis for making the decision of whether to adopt or discard a new experimental treatment, but additional methods are necessary for inference such as *P*-values, point estimates, and confidence intervals.

In a frequentist framework, *P*-values are commonly used to quantify the evidence for or against a hypothesis. In classical hypothesis testing, a properly computed *P*-value is uniformly distributed over the interval (0, 1) under the null hypothesis of no treatment effect. Computation of the *P*-value depends on the sampling scheme, however. For example, consider a score statistic $\mathcal{S} \dot{\sim} \mathcal{N}(\psi V, V)$ resulting from a fixed sample design. Under the null hypothesis, $\mathrm{H}_0 : \psi = 0$, the nominal two-sided *P*-value is computed as $p = 2(1 - \Phi(|z|))$, where $z = \mathcal{S}/(V)^{1/2}$ denotes the normalized *Z*-statistic. In the fixed sample setting, p is uniformly distributed over the interval (0, 1).

When data are gathered according to a group sequential design it is possible to compute the fixed sample *P*-value, defined as the nominal *P*-value corresponding to the test statistic obtained upon stopping. However, this is not a proper *P*-value as it is not distributed uniformly between 0 and 1 under the null hypothesis. To see this, consider a stopping rule with continuation sets C_j for $j = 1, \ldots, J$, and define the group sequential test statistic (M, S) as $M = \min\{1 \leq j \leq J : S_j \notin C_j\}$ and $S = S_M$, where S_M denotes the score statistic calculated at time M. For $\mathcal{S}_j \dot{\sim} N(\psi V_j, V_j)$, under the assumption of an independent increments structure the sampling density $p(m, \mathcal{S}_m; \psi)$ for the test statistic $(m, \mathcal{S}_m), m = 1, \ldots, J, \mathcal{S}_m \in (-\infty, \infty)$ is given by Armitage et al. (1969) as

$$p(m, \mathcal{S}_m; \psi_m) = \begin{cases} f(m, \mathcal{S}_m; \psi), & \mathcal{S}_m \notin \mathcal{C}_m, \\ 0, & \text{otherwise}, \end{cases} \quad (1)$$

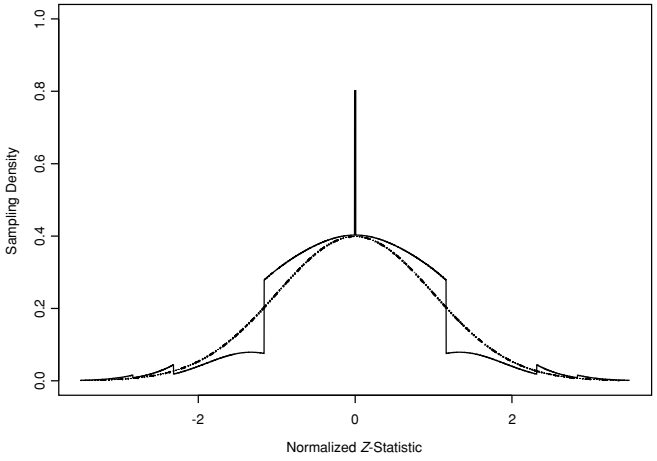
where the function $f(j, \mathcal{S}_j; \psi)$ is recursively defined as

$$f(1, \mathcal{S}_1; \psi) = \frac{1}{\sqrt{V_1}} \phi\left(\frac{\mathcal{S}_1 - \psi V_1}{\sqrt{V_1}}\right)$$

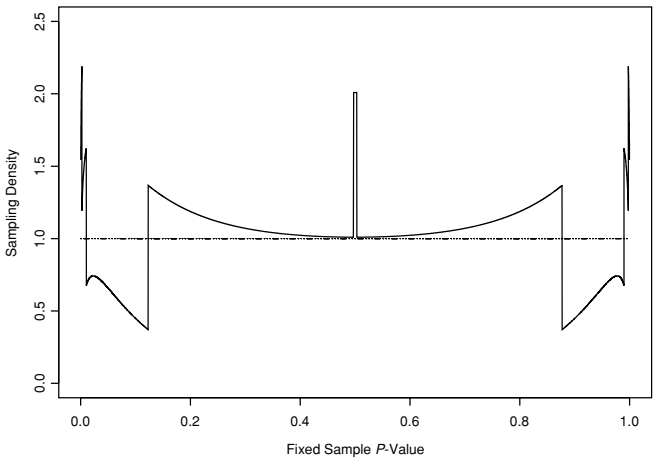
$$f(j, \mathcal{S}_j; \psi) = \int_{C_{j-1}} \frac{1}{\sqrt{v_j}} \phi\left(\frac{\mathcal{S}_j - s - \psi v_j}{\sqrt{v_j}}\right) f(j-1, s; \psi)\, ds, \qquad j = 2, \ldots, m$$

with $v_j = V_j - V_{j-1}$ for $j = 2, \ldots, m$, and $\phi(x) = e^{-x^2/2}/(2\pi)^{1/2}$ denoting the density of the standard normal distribution.

Figure 1 illustrates the effect of interim analyses on the distribution of the normalized *Z*-statistic and the fixed sample *P*-value under a level 0.05 two-sided symmetric stopping rule having O'Brien–Fleming boundary relationships across four equally spaced analyses (Emerson and Fleming, 1989). Under $H_0 : \psi = 0$ when data are gathered according to the group sequential design, the distribution of the *Z*-statistic is multimodal with jump discontinuities corresponding to the stopping boundaries at each of the analyses (Figure 1a; solid line). Subsequently, the fixed sample *P*-value is not uniformly distributed between 0 and 1 (Figure 1b; solid line). For reference, when data are gathered according to a fixed sample design the normalized *Z*-statistic follows a standard normal distribution and, as noted above, the fixed sample *P*-value follows a uniform distribution (Figure 1a and 1b, respectively; dashed lines).



(a)



(b)

Figure 1. Comparison of sampling densities for the normalized *Z*-statistic and fixed sample *P*-value statistic under $H_0 : \psi = 0$ when data are gathered according to a group sequential (———) and fixed sample (-----) design. The chosen group sequential design is a level 0.05 two-sided symmetric stopping rule having O'Brien–Fleming boundary relationships with four equally spaced analyses. (a) *Z*-statistic ($H_0 : \psi = 0$) and (b) *P*-value ($H_0 : \psi = 0$).

Various orderings of the bivariate statistic (M, S) have been proposed for the computation of proper *P*-values (adjusted for the stopping rule) that are uniformly distributed over the interval (0, 1), and several authors have suggested criteria by which these orderings should be judged. Emerson and Fleming (1990) ranked orderings by the degree of precision to which treatment effect parameters could be estimated. In particular, they compared the expected width of corresponding confidence intervals. Chang, Gould, and Snapinn (1995) compared power functions resulting from the analysis time (AT) (Tsiatis, Rosner, and Mehta, 1984), sample mean (Emerson and Fleming, 1990), *Z*-statistic (Chang, 1989), and score statistic (Rosner and Tsiatis, 1988) orderings. Recently, Cook (2002) compared these same orderings by considering the degree to which adjusted *P*-values agree with likelihood-based inference.

Emerson and Fleming (1990) found that the sample mean ordering provided uniformly shorter confidence intervals than the AT ordering and outperformed the *Z*-statistic ordering in some instances. Both Chang et al. (1995) and Cook (2002) concluded that the *Z*-statistic ordering is preferred under their respective criteria; however, Chang et al. (1995) found that the difference between the *Z*-statistic, AT, and sample mean orderings was relatively small for reasonably sized alternatives. These studies also agreed that the score statistic ordering performed poorly for alternatives sufficiently far from the null hypothesis. Finally, Jennison and Turnbull (2000) recommended the use of the AT ordering because (i) it ensures that adjusted *P*-values are less than the specified significance level if and only if the null hypothesis is rejected, and (ii) adjusted *P*-values do not condition on information levels beyond the stage at which the trial is stopped. Although (i) is guaranteed for all four of the orderings mentioned above, the AT ordering is the only one for which (ii) also holds.

Each of the authors noted above only considered these orderings in the case of time-invariant treatment effects, as with proportional hazards models. However, treatment effects on survival can frequently be delayed due to subgroup differences in which a certain portion of patients cannot be adequately treated or because the mechanistic path of the treatment may be long (see, for example, Abrams et al., 1994). Due to heavy dependence of the AT ordering on the timing of observed treatment effects, it is of interest to compare the power functions obtained under this ordering with those of the *Z*-statistic ordering which is independent of stopping time. We do not consider the sample mean ordering in this

manuscript because of our focus on nonproportional hazards: Because the log-rank statistic does not consistently estimate the same parameter over time in this setting, there is no well-defined estimate on which to base an ordering. In Section 2 we describe the framework used for comparing the AT and Z-statistic orderings and in Section 3 we contrast power functions obtained under these orderings for various nonproportional hazards configurations. Section 4 concludes with a brief discussion of the choice of outcome space orderings following a group sequential test and the implication of our results to clinical trials with longitudinal outcomes.

2. Comparison Framework

We consider the common clinical trial scenario of comparing the survival time of two groups, and assume the effect of treatment on the hazard for failure varies with time. In particular, we consider the settings of early and late diverging hazards, as depicted by the survival configurations in Figure 2. For illustration purposes it was assumed that baseline survival was distributed Exp(0.5) (median survival of approximately 1.4 years) in the early treatment effect setting and Exp(0.3) (median survival of approximately 2.3 years) in the delayed treatment effect setting. To reflect the staggered patient accrual rates typically encountered in clinical trials, patient entry times were taken to be uniformly distributed over 3 years with an additional follow-up of 1 year. Treatment effects were defined by two parameters: τ denoting the time at which treatment effects on the hazard scale start or stop, and θ denoting the log-hazard ratio comparing treatment to control. For each value of τ, sample sizes were adjusted to maintain 80% power for a fixed sample level 0.025 log-rank test when θ was assumed to be log(0.5).

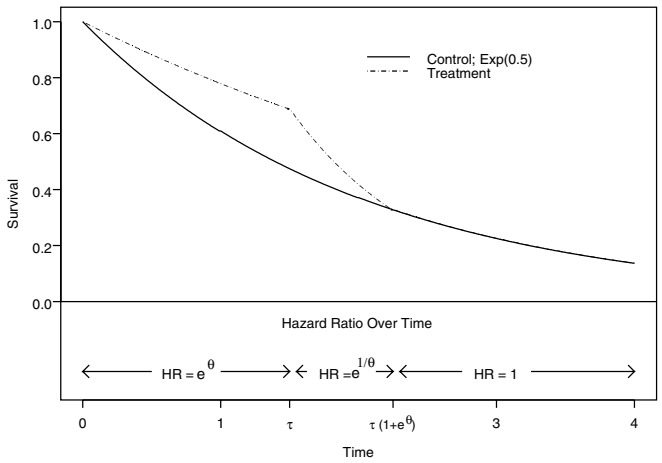


Figure 2. Survival configurations used to produce power functions in the early (a) and late (b) treatment effect settings.

In the current manuscript, we focus on postanalysis inference when P-value adjustment is based on the AT and Z-statistic orderings, defined as:

AT ordering (Tsiatis et al., 1984).

$$(M_1, S_1) < (M_2, S_2)$$

$$\text{iff } \begin{cases} M_1 < M_2 \text{ and } S_1 < x, & \forall x \in \mathcal{C}_{M_1}, \text{ or} \\ M_1 = M_2 \text{ and } S_1 < S_2, & \text{or} \\ M_1 > M_2 \text{ and } S_2 > x, & \forall x \in \mathcal{C}_{M_2}. \end{cases}$$

Z-statistic ordering (Chang, 1989).

$$(M_1, S_1) < (M_2, S_2) \quad \text{iff } z_1 = (S_1/\sqrt{V_1}) < (S_2/\sqrt{V_2}) = z_2.$$

The Z-statistic ordering was originally introduced by Chang (1989) as the likelihood ratio ordering, where it was used for the computation of confidence intervals following group sequential testing. Because we are presently only concerned with P-value calculations, it is sufficient to only consider the likelihood under the null hypothesis of no treatment effect, thereby basing our decision on the value of the Z-statistic.

Power functions, defined as pr$\{P\text{-value} \leq \alpha \mid \tau, \theta\}$ (Chang et al., 1995), were estimated under level 0.025 one-sided Pocock (1977) and O'Brien and Fleming (1979) stopping rules with four analyses. For each combination of τ and θ, the probability of obtaining a P-value $\leq \alpha \in (0, 0.05)$ under the AT and Z-statistic orderings was estimated based upon 25,000 simulations. Our interest in the power function is because the probability of observing, say $p < 0.001$, is of particular importance when a single clinical trial is to be used as a "pivotal study" for regulatory approval.

Without prior knowledge of a time-varying treatment effect, analyses are generally equally spaced in information time with the goal of balancing loss of statistical power against the potential for early stopping. In the event that one had a priori knowledge regarding a time-varying effect, it may be beneficial to shift analyses earlier or later in time. However, such knowledge is typically scarce, and the price paid for incorrectly assuming a particular treatment by time interaction may be a substantial loss in power or a significant increase in sample size. In the simulations presented here, we assumed no prior knowledge of a time-varying effect and adopted the common strategy of spacing analyses equally in information time.

Tests for survival differences were based upon the log-rank statistic. In practice, if one had scientific or clinical reason for believing survival differences at one time point were more

or less important than survival differences at other times, a weighted version of the log-rank statistic such as a member of the $G^{\rho,\gamma}$ family (Fleming and Harrington, 1991) may be more appropriate. With this said, we have chosen to focus on the (unweighted) log-rank statistic because such scientific knowledge is generally unknown at the start of a trial when the statistical protocol is to be defined. In this case, the log-rank statistic typically serves as a standard statistic for comparing survival distributions.

Although analyses were equally spaced in information time, for interpretation it is useful to consider the calendar times corresponding to each interim analysis. Under the assumption of Exp(λ) survival times, Unif(0, R) accrual, and testing based upon the log-rank statistic, the information fraction at calendar time t can be computed as (see, for example, Lan, Rosenberger, and Lachin, 1995)

$$\eta(t) = \begin{cases} \dfrac{\dfrac{t}{R} - \dfrac{1}{\lambda R}[1 - e^{-\lambda t}]}{1 + \dfrac{1}{\lambda R}[e^{-\lambda T} - e^{-\lambda(T-R)}]}, & t \le R, \\ \dfrac{1 + \dfrac{1}{\lambda R}[e^{-\lambda t} - e^{-\lambda(t-R)}]}{1 + \dfrac{1}{\lambda R}[e^{-\lambda T} - e^{-\lambda(T-R)}]}, & t > R, \end{cases} \tag{2}$$

where T is the maximal calendar duration of the trial. Thus, from equation (2) the calendar times of analyses in our comparison framework could be obtained under the null hypothesis of no treatment effect assuming $R = 3$ and $T = 4$ with four analyses equally spaced in information time, corresponding to $\eta(t) = 0.25$, 0.50, 0.75, and 1.0. For example, under the early treatment effect setting with baseline survival distributed Exp(0.5), interim analyses occurred at roughly $t =$ 19.6, 29.3, 37.6, and 48 months under the null hypothesis. Similarly, for the late treatment effect with baseline survival distributed Exp(0.3), interim analyses occurred at roughly $t =$ 20.9, 30.6, 38.6, and 48 months under the null hypothesis. Under nonproportional hazards alternatives, analyses were performed after $D/4$, $D/2$, $3D/4$, and D events had occurred, where D denotes the maximal number of planned events.

3. Power Functions under Nonproportional Hazards

By definition, under the AT ordering, once the decision to continue past an interim analysis is made, computed P-values are bounded from below by the probability, under the null hypothesis, of observing a partial sum statistic greater than or equal to the stopping boundary at that time. That is, the P-value will be bounded from below by the cumulative type I error spent over all preceding analyses. Table 1 displays the cumulative type I error spent for level 0.025 one-sided Pocock and O'Brien–Fleming stopping rules with four equally spaced analyses. Under the AT ordering, if treatment effects do not appear until the third analysis, the lowest P-value which could be obtained is 0.016 and 0.002 under the Pocock and O'Brien–Fleming designs, respectively. In contrast, because the Z-statistic ordering does not consider the analysis stage, there is no lower bound for P-values computed under this ordering, regardless of the stopping time.

Table 1
Information growth and error spent at each analysis for the one-sided Pocock and O'Brien–Fleming stopping rules

Analysis	Proportionate information	Error spent	
		Pocock	O'Brien–Fleming
1	0.25	0.009	<0.001
2	0.50	0.016	0.002
3	0.75	0.021	0.011
4	1.00	0.025	0.025

Table 2
Simulated probability of attaining a P*-value* $\le \alpha$

Treatment effect	Ordering	α			
		0.000625	0.001	0.01	0.025
Proportional hazards					
		Fixed sample			
	Both	0.359	0.402	0.687	0.802
		Pocock			
	Z-statistic	0.078	0.106	0.466	0.724
	Analysis time	0.021	0.036	0.226	0.717
		O'Brien–Fleming			
	Z-statistic	0.171	0.228	0.616	0.795
	Analysis time	0.089	0.117	0.530	0.792
Early effect					
		Fixed sample			
12 months	Both	0.367	0.41	0.694	0.804
18 months	Both	0.375	0.42	0.698	0.802
		Pocock			
12 months	Z-statistic	0.884	0.908	0.985	0.999
	Analysis time	0.891	0.915	0.996	0.999
18 months	Z-statistic	0.333	0.395	0.763	0.924
	Analysis time	0.244	0.291	0.787	0.924
		O'Brien–Fleming			
12 months	Z-statistic	0.979	0.984	0.992	0.992
	Analysis time	0.980	0.986	0.992	0.992
18 months	Z-statistic	0.646	0.700	0.861	0.884
	Analysis time	0.648	0.701	0.877	0.884
Delayed effect					
		Fixed sample			
6 months	Both	0.365	0.41	0.697	0.808
12 months	Both	0.379	0.426	0.702	0.801
		Pocock			
6 months	Z-statistic	0.096	0.135	0.472	0.696
	Analysis time	<0.001	0.001	0.031	0.685
12 months	Z-statistic	0.186	0.231	0.526	0.694
	Analysis time	<0.001	0.001	0.010	0.683
		O'Brien–Fleming			
6 months	Z-statistic	0.147	0.199	0.593	0.792
	Analysis time	0.013	0.018	0.376	0.789
12 months	Z-statistic	0.205	0.259	0.607	0.795
	Analysis time	0.001	0.001	0.168	0.792

Table 2 displays the simulated probability of obtaining a P-value $\le \alpha$, $\alpha = 0.000625$, 0.001, 0.01, 0.025, for $\theta =$ log(0.5) under proportional hazards, early, and delayed treatment effects. For reference, power functions under a level $\alpha = 0.025$ fixed sample test with power held constant at 80% are given. Note that P-values under the two orderings are equal for the fixed sample test. Under proportional hazards and early treatment effects ($\tau = 12$ and 18 months), the

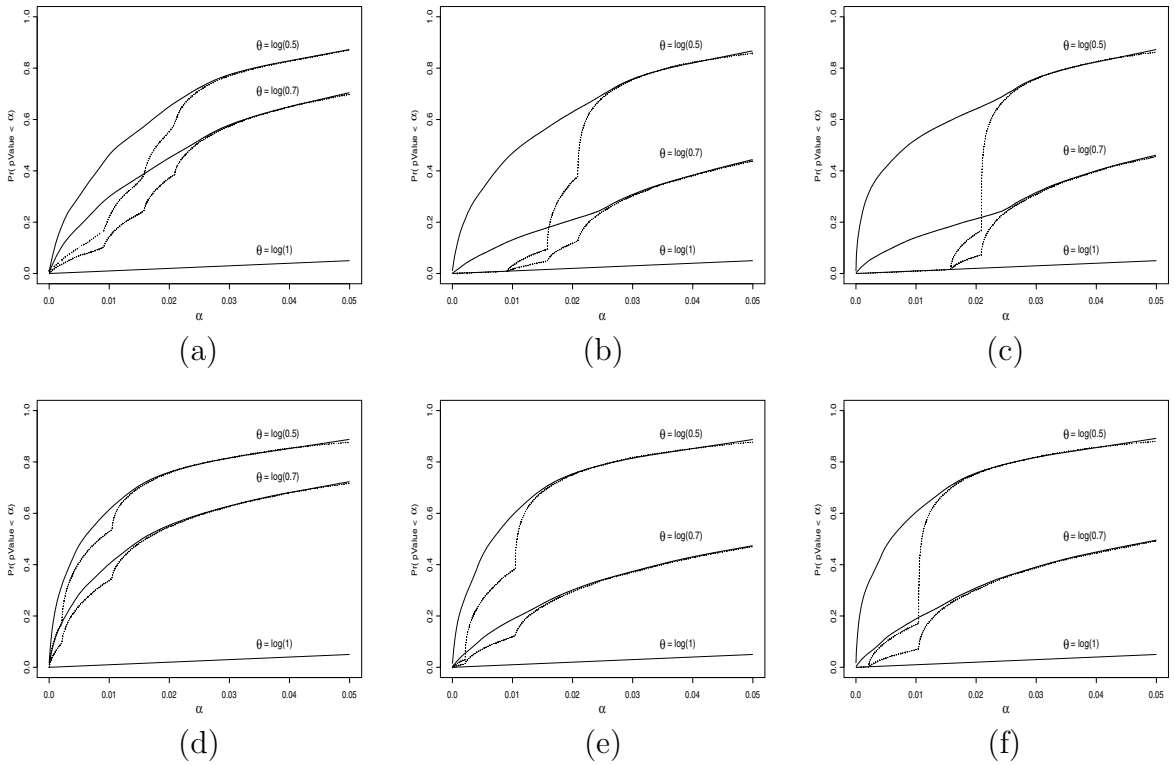


Figure 3. Simulated probability of attaining a P-value $\leq \alpha$ when the Z-statistic (———) and AT ordering (-----) are applied to data sampled under a time-delayed treatment effect. (a) Proportional hazards, (b) 6-month delay, and (c) 12-month delay were constructed under a Pocock stopping rule. (d) Proportional hazards, (e) 6-month delay, and (f) 12-month delay were constructed under an O'Brien–Fleming stopping rule.

AT and Z-statistic orderings result in comparable power estimates for both the Pocock and O'Brien–Fleming designs. As expected, both group sequential stopping rules attain much higher power than the fixed sample design when early treatment effects wane over time. Under delayed treatment effects, Table 2 reveals that power under the AT ordering for $\alpha < 0.025$ can be substantially lower than that based upon the Z-statistic ordering. For example, when treatment effects are delayed 12 months and an O'Brien–Fleming boundary is applied, the probability of obtaining a P-value < 0.01 under the Z-statistic ordering is 0.607 compared to 0.168 under the AT ordering. For the same treatment effect setting, using a Pocock boundary results in a probability of 0.526 for obtaining a P-value < 0.01 under the Z-statistic ordering compared to 0.010 under the AT ordering.

Figure 3 displays power curves for the Pocock and O'Brien–Fleming designs under proportional hazards and delayed treatment effects. Power functions under the two orderings are relatively comparable in the proportional hazards setting, although they do separate for large hazard differences in the case of the Pocock design. In contrast, under 6- and 12-month delayed treatment effects, the power function under the Z-statistic ordering easily dominates that under the AT ordering for values of $\alpha < 0.025$. The large discrepancy between power functions in this setting is attributable to performing interim analyses prior to the onset of treatment benefit, thus resulting in a delay in the rise of the AT power function.

While the power curves displayed here are specific to the schedule of analyses and test statistic we have chosen to use, they clearly demonstrate a potential drawback of the AT ordering. In general, if the analysis schedule is such that testing occurs prior to the initiation of treatment effects or if the chosen test statistic is unable to sufficiently detect late occurring treatment differences at the time of early analyses, the power function corresponding to P-values computed under the AT ordering will be adversely affected.

4. Discussion

Both the AT and Z-statistic orderings have been proposed for defining the extremity of data and are commonly used for computing proper P-values following a group sequential procedure. In the case of a constant treatment effect (e.g., proportional hazards), power functions under either the AT or Z-statistic ordering tend to be fairly comparable. Intuitively this is reasonable because large discrepancies in computed P-values can only appear in the event that a trial continues past early analyses and obtains an extremely large Z-statistic at a later analysis. Jennison and Turnbull (2000) recommend the use of the AT ordering because it does not require information regarding group sizes past the stopping stage and claim that the overshoot described above is unlikely. Clearly, if treatment effects are time invariant, the probability of such an overshoot is quite low. However, if one considers a delayed treatment effect as we have, it is possible that early analyses

will be performed prior to the time at which survival differences are revealed, while later analyses occurring after the initiation of treatment effect can detect large differences between the comparison groups.

The *Z*-statistic ordering is often preferred because it considers the magnitude of the estimate of treatment effect when defining extreme results. Because of this, no lower bound is placed upon *P*-values computed under the *Z*-statistic ordering and these *P*-values tend to agree largely with likelihood-based inference (Cook, 2002). Under a constant treatment effect the *Z*-statistic ordering tends to lead to slightly lower *P*-values when compared to the AT ordering, though the difference in the two orderings is not substantial under reasonably sized alternatives. However, when treatment effects are delayed, we have demonstrated that *P*-values calculated under the AT ordering can be substantially lower than those calculated under the *Z*-statistic ordering. For this reason, the *Z*-statistic ordering is recommended in group sequential settings where nonproportional hazards may be present.

Because the log-rank statistic is nonparametric, neither the sample mean ordering investigated by Emerson and Fleming (1990) nor their criteria of a good estimator, which are based upon the precision of parameter estimates, are directly applicable to this investigation. However, in the case of uncensored outcomes measured longitudinally, the potential for time-varying treatment effects does exist, and the sample mean ordering would be another possibility as results could be ordered by a single parameter estimate (e.g., an average slope over time). Under delayed treatment effects in the longitudinal setting, similar problems would certainly exist with the AT ordering, but further investigation of the performance of the *Z*-statistic and sample mean orderings is needed.

Acknowledgements

The authors would like to thank the associate editor and reviewer whose insightful comments and suggestions helped to improve the clarity and presentation of the article.

References

Abrams, D., Goldman, A., Launer, C., et al. (1994). A comparative trial of didanosine or zalcitabine after treatment with zidovudine in patients with human immunodeficiency virus infection. *New England Journal of Medicine* **330,** 657–662.

Armitage, P., McPherson, C. K., and Rowe, B. C. (1969). Repeated significance tests on accumulating data. *Journal of the Royal Statistical Society, Series A* **132,** 235–244.

Chang, M. N. (1989). Confidence intervals for a normal mean following a group sequential test. *Biometrics* **45,** 247–254.

Chang, M. N., Gould, A. L., and Snapinn, S. M. (1995). *p*-values for group sequential testing. *Biometrika* **82,** 650–654.

Cook, T. D. (2002). *p*-value adjustment in sequential clinical trials. *Biometrics* **58,** 1005–1011.

Emerson, S. S. and Fleming, T. R. (1989). Symmetric group sequential test designs. *Biometrics* **45,** 905–923.

Emerson, S. S. and Fleming, T. R. (1990). Parameter estimation following group sequential hypothesis testing. *Biometrika* **77,** 875–892.

Fleming, T. R. and Harrington, D. P. (1991). *Counting Processes and Survival Analysis*. New York: Wiley.

Jennison, C. and Turnbull, B. W. (2000). *Group Sequential Methods with Applications to Clinical Trials*. Boca Raton, Florida: CRC Press.

Lan, K. K. G., Rosenberger, W. F., and Lachin, J. M. (1995). Sequential monitoring of survival data with the Wilcoxon statistic. *Biometrics* **51,** 1175–1183.

Mantel, N. (1966). Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemotherapy Reports* **50,** 163–170.

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

Pocock, S. J. (1977). Group sequential methods in the design and analysis of clinical trials. *Biometrika* **64,** 191–200.

Rosner, G. L. and Tsiatis, A. A. (1988). Exact confidence intervals following a group sequential trial: A comparison of methods. *Biometrika* **75,** 723–729.

Tsiatis, A. A. (1982). Repeated significance testing for a general class of statistics used in censored survival analysis. *Journal of the American Statistical Association* **77,** 855–861.

Tsiatis, A. A., Rosner, G. L., and Mehta, C. R. (1984). Exact confidence intervals following a group sequential test. *Biometrics* **40,** 797–803.

Received March 2004. *Revised July* 2004.
Accepted August 2004.