Flexible Weighted Log-Rank Tests Optimal for Detecting Early and/or Late Survival Differences

Lang Wu

Department of Statistics, University of British Columbia, Vancouver, B.C., V6T 1Z2, Canada

email: lang@stat.ubc.ca

and

Peter B. Gilbert

Fred Hutchinson Cancer Research Center and Department of Biostatistics,

University of Washington, Seattle, WA 98109, U.S.A.

SUMMARY. At the present time many AIDS clinical trials compare drug therapies by a time to event primary endpoint that measures the durability of suppression of HIV replication. For such studies, survival differences tend to occur early and/or late in the follow-up period, due to drug differences in initial potency and/or durability of efficacy, and detecting these differences is of primary interest. We propose a weighted log-rank statistic which emphasizes early and/or late survival differences. We also consider some versatile tests that also emphasize these differences, but are sensitive to a wider range of alternatives. The performances of the new tests are evaluated in numerical studies. For the alternatives of interest, the new tests show greater power and flexibility than commonly used weighted log-rank tests and related versatile tests. When the main interest is in detecting early and/or late survival differences, these tests may be preferable to the other versatile and weighted log-rank tests that have been studied.

KEY WORDS: AIDS; Asymptotics; Quadratic weight; Robust test; Versatile test.

1 Introduction

A primary goal of antiretroviral therapy for HIV infected individuals is to durably suppress HIV replication (Guidelines, 1998). With this aim in mind, many AIDS clinical trials compare treatments primarily by a time to "virologic failure" endpoint that signifies loss of viral suppression. Commonly this endpoint is defined by either a lack of initial suppression of plasma virus levels below a threshold during the early follow-up period (typically 8-16 weeks), or by a confirmed rise in plasma virus levels above a threshold in subjects who achieved initial virologic suppression (Gilbert et al., 2000). Often a relatively high failure rate is observed during the early period, driven by impotency of the therapies for some subjects and/or by pre-existing drug-resistant viruses. Subjects whose virus is successfully suppressed during the early period typically experience sustained suppression for many months, with event rates increasing late in the follow-up period, due to emergence of drug toxicity and/or drug resistance (Guidelines, 1998). Because the therapies tend to fail early and/or late in the follow-up period, it is of interest to compare the treatment arms by a testing procedure that is sensitive to detect early and/or late treatment differences. Furthermore, since the relative degree to which the therapies are expected to differ by early or late differences is usually unknown, it may be prudent to take an approach that is sensitive for detecting both kinds of differences. In this article, such testing procedures are developed.

AIDS Clinical Trials Group (ACTG) Study 364 provides an example of a clinical trial where early and late treatment differences occurred in the survival curve of virologic failure (see Figure 1). This randomized trial evaluated the ability of three combination drug regimens to durably suppress HIV replication, that included the drugs (1) efavirenz (EFV), (2) nelfinavir (NFV), and (3) EFV combined with NFV. Figure 1 shows point estimates and 95% pointwise and simultaneous confidence intervals for the log hazard ratio over the time period 5 to 80 weeks, for the EFV arm versus the EFV + NFV arm. The estimates were computed by nonparametric kernel smoothing of the group-specific hazard functions, with confidence intervals calculated by the simulation technique of Gilbert et al. (2001). The plot exhibits superiority of EFV + NFV in the early and late periods, with modest or absent treatment differences in the middle period. This suggests that the regimen containing more drugs (EFV + NFV) provides a more potent initial treatment, and more sustained efficacy in those with responsive virus. To detect such early and/or late survival differences, it is expected that test statistics that assign more weight on early and late survival differences will be more sensitive than test statistics that weight all survival differences equally, such as the log-rank statistic, or than test statistics that weight only early differences or only late differences, such as weighted log-rank statistics described in the literature.

Weighted log-rank tests have been widely used for testing the equality of two survival distributions in the presence of arbitrary right censoring. These tests may give more weight on either early, middle, or late survival differences (Fleming and Harrington, 1991). A particularly useful class is the so-called G^ρ family. This class contains as special cases the well-known log-rank statistic (ρ=0), which assigns equal weights throughout the study, and the generalized (Peto-Peto) Wilcoxon statistic ($\rho = 1$), which assigns more weight on early survival differences. It is known that in many situations, properly choosing a weight function can lead to a substantial increase in efficiency, and it is obvious that test statistics that give greater (lesser) weight to events occurring early in the study will be more (less) sensitive to early differences in survival distributions. In comparing the potency and/or durability of treatment responses in AIDS studies, we are mainly interested in testing early *and/or* late survival differences. Thus, in the present paper we propose a weighted log-rank test statistic that gives more weight to early and late events. The proposed test is expected to be more efficient than existing weighted log-rank tests in detecting early and late survival differences simultaneously, and more robust when one wishes to detect either kind of difference.

If the shape of the log hazard ratio as a function of follow-up time is known, optimal weighted log-rank tests can be derived by incorporating this shape into the weight function (Fleming and Harrington, 1991). Our proposed test uses a quadratic weight function that gives more weight to early and late events, so that it will be sensitive for alternatives where the log hazard ratio exhibits quadratic forms. As shown in Figure 1, the estimated log hazard ratio for two treatments in ACTG Study 364 has a nearly quadratic from. Such quadratic log hazard ratio alternatives have also been observed in other studies (e.g., Stablein et al., 1981).

In many situations, it may not be possible to specify in advance the types of survival differences that will actually occur. Misspecification of the weights in weighted log-rank test statistics can have undesirable consequences for the power of the tests. Some versatile testing procedures, such as the supremum versions of the linear rank statistics, have been proposed (Fleming et al., 1987; Self, 1991). Such versatile tests have been shown to provide good power against a wide range of alternatives. Recently, Shen and Cai (2001) proposed a versatile test procedure based on a maximum of a set of non-rank-based, weighted test statistics. For their versatile test, the individual test statistics in the set are weighted Kaplan-Meier statistics, which may be more sensitive to alternatives with a lag time and more sensitive to the magnitude of the survival difference than the rank-based statistics. They have mainly focused on cancer prevention and screening trials with early, or late, or middle survival differences. In this article, we propose some new versatile tests based on our proposed quadratic weighted log-rank statistic, that are of particular interest in AIDS therapeutic studies.

In Section 2, we describe our proposed weighted log-rank statistic and study its asymptotic properties. In Section 3, we consider some versatile tests. The performances of the newly proposed tests are evaluated in Section 4 by simulations. Section 5 applies the tests to the ACTG Study 364 example. Some issues are discussed in Section 6.

2 A Linear Rank Test for Early and Late Differences

2.1 *The New Test*

Consider a clinical trial of duration τ with two treatment arms. Let n_i $(i = 1, 2)$ be the number of subjects in treatment arm i and let $n = n_1 + n_2$. Assume that the observations from the n subjects are mutually independent. Let X_{ij} and Y_{ij} be the failure times and censoring times for subject j in treatment arm i respectively, $i = 1, 2;\ j = 1, \cdots, n_i$. The distribution functions of X_{ij} and Y_{ij} are denoted by F_i and G_i, respectively. For simplicity, we assume that F_i and G_i are continuous. We also assume that X_{ij} and Y_{ij} are independent. Let $S_i(t) = Pr(X_{ij} > t) = 1 - F_i(t)$ and $C_i(t) = Pr(Y_{ij} > t) = 1 - G_i(t)$. Let $\widehat{S}_i(t)$ and $\widehat{C}_i(t)$ be the left continuous versions of the Kaplan-Meier estimators of $S_i(t)$ and $C_i(t)$, respectively, and let $\widehat{S}(t)$ be the left continuous version of the Kaplan-Meier estimator of $S(t)$ in the pooled sample.

The observable random variables are $T_{ij} = \min(X_{ij}, Y_{ij})$ and $\delta_{ij} = I(X_{ij} \leq Y_{ij})$, where $I(A)$ is 1 if the event A occurs and 0 otherwise. Let $Y_i(t) = \sum_{j=1}^{n_i} I(T_{ij} \geq t)$ be the size of the risk set at time t in sample i, and let $N_i(t) = \sum_{j=1}^{n_i} I(T_{ij} \leq t, \delta_{ij} = 1)$ be the number of observed failures in sample i that have occurred by time t. The null hypothesis of interest is $H_0 : S_1 = S_2$.

The general form of a weighted log-rank statistic, when based on information over 0 to t units of survival time, can be expressed as

$$T(W, t) = \int_0^t W(s) \left\{ \frac{Y_1(s) Y_2(s)}{Y_1(s) + Y_2(s)} \right\} \left\{ \frac{dN_1(s)}{Y_1(s)} - \frac{dN_2(s)}{Y_2(s)} \right\}, \tag{1}$$

where $W(s), s \geq 0$, is a weight function. For the G^ρ family, $W(s) = \widehat{S}^\rho(s-)$ with $0 \leq \rho < \infty$. Note that, for $\rho > 0$, $\widehat{S}^\rho$ is a decreasing function; thus the G^ρ $(\rho > 0)$ statistics give more weight to early survival differences, while for $\rho = 0$, the log-rank statistic G^0 weights survival differences equally throughout the study. When the main interest is in detecting differences in cure rates, Gray and Tsiatis (1989) proposed a statistic which has the same form as the G^ρ family but with $\rho = -1$. For censored data, their test can be used to test the equality of survival distributions with power focused against late differences. They showed that their test is optimal under the proportional distributions alternative.

In AIDS treatment trials, we are interested in early and/or late survival differences, not merely early differences or merely late differences. Since we often do not know whether the differences are due to early events or late events or both, it is desirable to give more weight on both early and late survival differences. We thus propose the following weighted log-rank statistic based on a simple quadratic weight function:

$$\widetilde{G}_a^2 \equiv \int_0^\tau \left\{\widehat{S}(s-) - (a\widehat{S}(\tau-) + 1 - a)\right\}^2 \left\{\frac{Y_1(s)Y_2(s)}{Y_1(s)+Y_2(s)}\right\} \left\{\frac{dN_1(s)}{Y_1(s)} - \frac{dN_2(s)}{Y_2(s)}\right\}, \qquad (2)$$

where a is a known constant such that $0 \le a \le 1$. The proposed quadratic weight function $W_a(s) \equiv \left\{\widehat{S}(s-) - (a\widehat{S}(\tau-) + 1 - a)\right\}^2$, when scaled by $a^2(1-\widehat{S}(\tau-))^2$ for $a > 0$ and $\widehat{S}(\tau-) < 1$, decreases from 1 to 0 and then increases from 0 to $a^2/(1-a)^2$, and attains 0 (minimum) weight roughly when $a\times 100\%$ of the observed failures occur. The ratio of the weights given at time $t = 0$ and time $t = \tau$ is $W_a(0)/W_a(\tau) = (a/(1-a))^2$, which ranges from 0 to ∞ as a varies from 0 to 1. For example, if $a = 0.5$, the proposed test statistic $\widetilde{G}_{0.5}^2$ weights early and late events equally and gives minimal weight to the "middle time" when 50% of the observed failures occur. For $a = 1$ and $\widehat{S}(\tau-) = 0$, $\widetilde{G}_a^2$ reduces to the G^ρ family with $\rho = 2$ which emphasizes early differences, and for $a = 0$ and $\widehat{S}(\tau-) = 0$, $\widetilde{G}_a^2$ reduces to the $G^{\rho,\gamma}$ family with $\rho = 0$ and $\gamma = 2$ which emphasizes late differences (Fleming and Harrington, 1991).

The ratio $W_a(0)/W_a(\tau)$ indicates that, by choosing the appropriate value of a in $\widetilde{G}_a^2$, we can weight early and late events differently at any desired proportion. Sometimes, prior experience can help us to determine the value of a in advance that gives the desired weight function. Figure 2 shows the shape of the weight function $W_a(s)$ for various values of a.

Following similar arguments as in Fleming and Harrington (1991), a natural estimator of the variance of $\widetilde{G}_a^2$ is given by $\widehat{\sigma}^2_{\widetilde{G}_a^2} = \int_0^\tau W_a(s)^2\{Y_1(s)Y_2(s)/(Y_1(s)+Y_2(s))^2\}\{dN_1(s)+dN_2(s)\}$, and, under H_0, we have $\widetilde{G}_a^2/\widehat{\sigma}_{\widetilde{G}_a^2} \to N(0,1)$. Note that, strictly speaking, the new weight function is not predictable with respect to the natural σ-filtration generated by the counting processes $\{N_i(t), Y_i(t)\}$. Based on an argument similar to that in Buyske et al. (2000), however, standard martingale results are still valid (see Appendix).

2.2 *Asymptotic Properties of the New Test Statistic*

It is known that the G^ρ statistics are asymptotically efficient under sequences of local alternatives generated from certain semiparametric transformation models (Fleming and Harrington, 1991). In this section we discuss the asymptotic properties of our new test statistic $\widetilde{G}_a^2$, and derive sequences of local (contiguous) alternatives where the $\widetilde{G}_a^2$ statistic is asymptotically efficient. The contiguous alternative may be used to determine the appropriate sample size for any given study.

Following Fleming and Harrington (1991) and Buyske et al. (2000), we consider the contiguous alternative: $f_2(t) = f_1(t) + n^{-1/2}b(t)f_1(t) + o(n^{-1/2})$, where $f_i(t) = F_i'(t)$ and $b(t)$ is a continuous function on $[0,\tau]$. Suppose that $n_i/n \to r_i,\ i = 1,2$. Under this alternative, it can be shown by standard methods that (see Appendix) $\widetilde{G}_a^2/\widehat{\sigma}_{\widetilde{G}_a^2} \to N(\mu_a, 1)$, where $\mu_a = \int_0^\tau W_a^*(t)b(t)d\mu(t)/\left(\int_0^\tau W_a^*(t)^2 d\mu(t)\right)^{1/2}$, $d\mu(t) = (r_1C_1(t)r_2C_2(t)/r_1C_1(t)+r_2C_2(t))dF_1(t)$, and $W_a^*(t)$ is the non-random limit of $W_a(t)$ (see Appendix). Thus the asymptotic power is $1-\Phi(z_\alpha-\mu_a)$, where Φ and z_α are the distribution function and the $(1-\alpha)\times 100$th percentile of the standard normal distribution, respectively. Assuming $\mu_a > 0$, the one-sided level α test rejects H_0 if $\widetilde{G}_a^2/\widehat{\sigma}_{\widetilde{G}_a^2} > z_\alpha$. Moreover, under the above contiguous alternative and the semiparametric transformation model (Fleming and Harrington, 1991) $F_1(t) = H(g(t)),\quad F_2(t) = H(g(t)+\theta)$, where H and g are monotone increasing functions and θ is the parameter indicating treatment effect, the statistic $\widetilde{G}^2_a$ is asymptotically optimal if $H(u) = aa_2 - L^{-1}_{a,a_2}(a_1(u^* - u))$ for some $a_1 > 0$ and $a_2 > 0$, where $L_{a,a_2}(v) = \int_{(a-1)a_2}^{v} ((aa_2)^3 - x^3)^{-1} dx$, $(a-1)a_2 \leq v \leq aa_2$, and $u = g(t)$ and $u^* = g(\tau)$. The proof of this result is similar to that of Theorem 4 of Buyske et al. (2000), so is omitted here.

3 Some Versatile Tests

3.1 *A Supremum Statistic*

As noted in Section 2, the G^ρ statistics have good power to detect proportional hazards alternatives, early survival differences, and late survival differences. We will show in Section 4 that our proposed new statistic $\widetilde{G}^2_a$ has good power to detect early and/or late differences. All these procedures, however, can have low power against other important, realistic alternatives, as will be seen later. Since in practice we may not know the type of alternatives in advance, it is important to study statistics that will be sensitive to a wider range of alternatives. Fleming et al. (1987) investigated supremum versions of the statistics G^0 and G^1, which are more versatile than their corresponding statistics. In particular, they showed that their κ^0 and κ^1 statistics are the best among other supremum versions of the weighted log-rank statistics. Self (1989) studied a class of versatile statistics which essentially contains supremum versions of the statistic $G^{\rho,\gamma}$ introduced by Fleming and Harrington (1991), with the supremum taken with respect to ρ and γ over a range of values.

As noted earlier, by choosing the appropriate value of a, the new statistic $\widetilde{G}^2_a$ will be sensitive for detecting early and late survival differences. In practice, however, the optimal value of a is often unknown. For a unknown, we propose the following supremum version of the quadratic statistic $\widetilde{G}_a^2$ which identifies the weights in a data-dependent way:

$$T(\mathcal{A}) = \ \sup_{a\in\mathcal{A}} \left[\int_0^\tau W_a(s) \left\{ \frac{Y_1(s)Y_2(s)}{Y_1(s)+Y_2(s)} \right\} \left\{ \frac{dN_1(s)}{Y_1(s)} - \frac{dN_2(s)}{Y_2(s)} \right\} \Big/ \widehat{\sigma}_{\widetilde{G}_a^2} \right], \tag{3}$$

where $\mathcal{A}$ is a subset of the interval $[0,1]$ which covers all reasonable values of a in a particular study (the most general case is $\mathcal{A} = [0,1]$). The adaptive statistic $T(\mathcal{A})$ is expected to be more versatile than any particular statistic $\widetilde{G}_a^2$ and at the same time be nearly as sensitive as the most powerful individual statistic. We will investigate its performance in Section 4.

The computation of the supremum $T(\mathcal{A}) = \sup_{a\in\mathcal{A}} \left\{ \widetilde{G}_a^2/\widehat{\sigma}_{\widetilde{G}_a^2} \right\} \equiv \sup_{a\in\mathcal{A}} h(a)$ can be achieved by the standard Newton-Raphson algorithm. To find a good starting value for a, we may first choose a finite set $\mathcal{A}_0 = \{a_1^*, \ldots, a_k^*\} \subset \mathcal{A}$ and compute $h(a) \equiv \widetilde{G}_a^2/\widehat{\sigma}_{\widetilde{G}_a^2}$ for each $a_i \in \mathcal{A}_0$. Then the $a_i^* \in \mathcal{A}_0$ which maximizes $h(a)$ on the finite set $\mathcal{A}_0$ can be taken as a starting value for the Newton-Raphson algorithm to find the overall supremum $T(\mathcal{A})$. Significance tests based on $T(\mathcal{A})$ can be obtained by bootstrap methods. For example, we use the following method, suggested in Davison and Hinkley (1997, Section 3.5.2). Draw a sample of size n_1 with replacement from the observed data pairs $\{(X_i, \delta_i)\}$, pooled over both groups, and draw a sample of size n_2 from the same data pairs. The two samples form a bootstrap dataset under $H_0 : S_1 = S_2$. For each of B re-sampled datasets, compute the test statistic $T(\mathcal{A})$, which gives B bootstrap test statistics $T_1^*(\mathcal{A}), \cdots, \mathcal{T}_{\mathcal{B}}^*(\mathcal{A})$. Then, the desired p-value is calculated as the proportion of the B bootstrap test statistics that exceed or equal the statistic $T(\mathcal{A})$ computed based on the originally observed data.

Note that two statistics $\widetilde{G}_a^2$ and $\widetilde{G}_b^2$ are highly correlated if a is very close to b. The null correlation between $\widetilde{G}_a^2$ and $\widetilde{G}_b^2$ can be computed using a formula given in Fleming and Harrington (1991). For example, in the uncensored case the correlation of $\widetilde{G}_{0.4}^2$ and $\widetilde{G}_{0.5}^2$ is 0.88, while the correlation of $\widetilde{G}_0^2$ and $\widetilde{G}_1^2$ is 0.17. Thus, in practice it is reasonable to only consider the cases where $\mathcal{A}$ is a finite set of plausible values a_i in $[0,1]$, such as those based on prior experience. When $\mathcal{A}$ is finite, say $\mathcal{A} = \{a_1, \cdots, a_k\}$ $(a_1 < a_2 < \ldots < a_k)$, $T(\mathcal{A})$ is simply the maximum of $\widetilde{T}_i \equiv \widetilde{G}^2_{a_i}/\widehat{\sigma}_{\widetilde{G}^2_{a_i}}$, $i = 1, \cdots, k$. Using the large-sample normality result (Fleming and Harrington, 1991), the joint asymptotic distribution of $(\widetilde{G}^2_{a_1}, \cdots, \widetilde{G}^2_{a_k})$ is multivariate normal with mean zero. Thus, we can calculate $P(T(\mathcal{A}) \geq t)$ for any t based on a program for computing multivariate normal probabilities (e.g., Genz, 1992), and one-sided p-values of the proposed test can be easily obtained. A two-sided test can be performed in a similar way.

3.2 *Linear Combination Statistics*

As an alternative to the supremum statistic $T(\mathcal{A})$, here we consider statistics based on linear combinations of the individual statistics $\widetilde{T}_i \equiv \widetilde{G}^2_{a_i}/\widehat{\sigma}_{\widetilde{G}^2_{a_i}}$: $\widetilde{T}(\mathbf{c}) \equiv \sum_{i=1}^k c_i \widetilde{T}_i$, where $c_i \geq 0$, $\sum_{i=1}^k c_i = 1$, and $\mathbf{c} = (c_1, \ldots, c_k)$ is a constant vector. We are interested in choosing $\mathbf{c}$ such that $\widetilde{T}(\mathbf{c})$ has reasonable efficiency relative to the optimal $\widetilde{T}_i$ (say) while being more robust than the individual statistics across the range of alternatives.

For two weighted log-rank statistics $\widetilde{T}_i$ and $\widetilde{T}_j$, it is known that the Pitman asymptotic relative efficiency (ARE) of $\widetilde{T}_i$ with respect to $\widetilde{T}_j$ is (Fleming and Harrington, 1991) $ARE(\widetilde{T}_i, \widetilde{T}_j) = \widehat{\sigma}_{ij}/(\widehat{\sigma}_{ii}\widehat{\sigma}_{jj})$, which is the square of the null correlation between $\widetilde{T}_i$ and $\widetilde{T}_j$. Let $\mathcal{C} = \{\mathbf{c} = (c_1, \ldots, c_k) : c_i \geq 0, \sum_{i=1}^k c_i = 1\}$, and let $\widetilde{T}(\mathcal{C}) = \{\widetilde{T}(\mathbf{c}) : \mathbf{c} \in \mathcal{C}\}$ be the class of all linear combination statistics. Following the idea of Gastwirth (1966) and Zucker and Lakatos (1990), we consider the *maximum efficiency robust test* (MERT) for $\widetilde{T}(\mathcal{C})$, denoted by $\widetilde{T}^* \equiv \widetilde{T}(\mathbf{c}^*)$, which maximizes $\inf_{1 \leq i \leq k} ARE(\widetilde{T}(\mathbf{c}), \widetilde{T}_i)$ over all $\widetilde{T}(\mathbf{c}) \in \widetilde{T}(\mathcal{C})$. In other words, among all linear combination statistics $\widetilde{T}(\mathbf{c})$, the MERT $\widetilde{T}^*$ minimizes the worst possible efficiency loss, over all $\{\widetilde{T}_i, 1 \leq i \leq k\}$, associated with using $\widetilde{T}(\mathbf{c})$ instead of the optimal statistic $\widetilde{T}_i$. Let $R = (\rho_{ij})_{k \times k}$ be the correlation matrix for $\{\widetilde{T}_i, 1 \leq i \leq k\}$. Following similar arguments as in Gastwirth (1966), the MERT for $\widetilde{T}(\mathcal{C})$ is given by $\widetilde{T}^* = \sum_{i=1}^k c_i^* \widetilde{T}_i$, where $\mathbf{c}^* = (c_1^*, \ldots, c_k^*)$

minimizes $\mathbf{c}^T R\mathbf{c}$, subject to $R\mathbf{c} \geq 1$ and $\mathbf{c} > 0$. The solution is given by $\mathbf{c}^* = R^{-1}\mathbf{e}$, where $\mathbf{e} = (1, \ldots, 1)^T$ (Luenberger, 1984). Another robust linear combination statistic, which may be considered as an approximation to $\widetilde{T}^*$ (Gastwirth, 1966) and may also be useful, is the sum of the two extreme pair given by $\widetilde{T}_0^* \equiv \widetilde{T}_1 + \widetilde{T}_k = \widetilde{G}^2_{a_1}/\widehat{\sigma}_{\widetilde{G}^2_{a_1}} + \widetilde{G}^2_{a_k}/\widehat{\sigma}_{\widetilde{G}^2_{a_k}}$. Under H_0, the linear combination statistics asymptotically follow univariate normal distributions with zero mean, and tests based on these statistics can be derived directly.

Note that other linear combination statistics can also be proposed. For example, in AIDS studies, if a significant difference is found based on $\widetilde{G}^2_a$, a linear combination statistic such as $G^1 - G^{-1}$ can be used as a subsequent test to assess whether the difference is due to early differences or late differences. In addition, the statistic $\widetilde{G}^2_a - (G^1 + G^{-1})/2$ can be used to test if the difference is due to early and late differences, and not only to early or late differences, and the statistics $\widetilde{G}^2_a - G^0, \widetilde{G}^2_a - G^1$ and $\widetilde{G}^2_a - G^{-1}$ may also be useful.

4 Simulations

To examine and compare the small and moderate sample behavior of the proposed statistics, we performed extensive Monte Carlo simulations. In the simulations, a null case and four distinct alternative cases were considered (Figure 3). These four survival configurations represent proportional hazards, early differences, late differences, and early and late differences. Some configurations are similar to those considered in Fleming et al. (1987). Data were generated from piece-wise exponential distributions. Data were either 20% censored or 60% censored. In order to mimic real situations, the censoring was created such that 90% were due to truncation (the truncation time was $\tau = 2$) and the remaining 10% were uniform over (0,2). To achieve the desired censoring rates and censoring mechanism, the parameter values of the alternatives for the 20% and 60% censored cases are different. Thus, the powers for the

20% censored case and the 60% censored case in each configuration are not comparable (i.e., they are not necessary decreasing). In each configuration, two sample sizes were considered: $n_1 = n_2 = 20$ and $n_1 = n_2 = 50$, and 1000 simulations were performed using a one-sided test with an $\alpha = 0.05$ significance level. The test statistics included in the simulations are G^0, G^1, G^{-1}, $\widetilde{G}_0^2$, $\widetilde{G}_1^2$, $\widetilde{G}_{0.5}^2$, $\widetilde{G}_{0.4}^2$, and the versatile tests $\kappa^0, \kappa^1, T(\mathcal{A})$, $\widetilde{T}_0^*$, and $\widetilde{T}^*$, where $\mathcal{A} = \{0.1, 0.2, \ldots, 0.9\}$. We also include a versatile test based on the simultaneous use of the four statistics G^0, G^1, G^{-1}, and $\widetilde{G}_{0.5}^2$: $G_{max} = \max\{G^0/\widehat{\sigma}_{G^0},\ G^1/\widehat{\sigma}_{G^1},\ G^{-1}/\widehat{\sigma}_{G^{-1}},\ \widetilde{G}_{0.5}^2/\widehat{\sigma}_{\widetilde{G}_{0.5}^2}\}$. P-values for this test are obtained using the joint asymptotic multivariate normal distribution of $(G^0, G^1, G^{-1}, \widetilde{G}_{0.5}^2)$.

Table 1 presents the simulation results on the sizes and powers of the various statistics. All tests achieve approximately correct sizes. As expected, G^0 is quite sensitive against the proportional hazards alternatives (configuration I), G^1(or $\widetilde{G}_1^2$) and G^{-1}(or $\widetilde{G}_0^2$) are quite sensitive against early (configuration II) and late (configuration III) survival differences respectively, and $\widetilde{G}_{0.5}^2$ (or $\widetilde{G}_{0.4}^2$) have good power against early and late survival differences (configuration IV). For the particular survival configurations considered in this simulation, the statistics $\widetilde{G}_1^2$ ($\widetilde{G}_0^2$) appear to be more powerful than G^1 (G^{-1}) in detecting early (late) survival differences. To detect early *and* late survival differences, the newly proposed statistics $\widetilde{G}_{0.5}^2$ (or $\widetilde{G}_{0.4}^2$) were demonstrated to have much greater power than other statistics ($\widetilde{G}_{0.4}^2$ emphasizes more late differences), and in the meantime have reasonable power in detecting early *or* late survival differences.

Among the versatile tests, the newly proposed statistics $T(\mathcal{A})$ and $\widetilde{T}^*$ appear to be preferable to $\widetilde{T}_0^*$, G_{max}, κ_0, and κ_1. They are more versatile than any other statistic while being nearly as sensitive as the most powerful individual statistic against each configuration. The κ_0 and κ_1 statistics, while being more versatile than and being nearly as sensitive as their corresponding non-supremum statistics G^0 and G^1, appear to have lower powers against late differences or early and late differences, as expected. Since in practice we often do not know the type of alternatives in advance, we also list the average powers of the tests over all four alternatives, and we again find that the newly proposed statistics $\widetilde{G}_a^2$, $T(\mathcal{A})$, $\widetilde{T}^*$, $\widetilde{T}_0^*$, and G_{max} perform well, and the new statistics $T(\mathcal{A})$ and $\widetilde{T}^*$ continue to be the best among all statistics.

To check the sensitivity of $T(\mathcal{A})$ and $\widetilde{T}^*$ to the selection of values in $\mathcal{A}$, we performed a simulation with $\mathcal{A}_1 = \{0.05, 0.1, 0.15, \ldots, 0.9, 0.95\} \supset \mathcal{A}$ and found very similar results to those in Table 1 (not shown). This probably occurred because the test statistics based on $\mathcal{A}$ are highly correlated, as noted in Section 3.1. Thus, when the set of test statistics are strongly correlated with one another, use of a larger set may not enhance power but may increase computational burden substantially.

5 Example

We now apply several of the tests evaluated in the simulations to compare the survival curves of virologic failure for the EFV and EFV + NFV arms that were studied in ACTG Study 364. The numbers of subjects randomized for the EFV and EFV + NFV arms were 65 and 64 respectively. We define the virologic failure endpoint as either "early virologic failure" within the first 16 weeks since randomization, or as two consecutive HIV-1 RNA measurements ≥ 200 copies/ml at week 16 or later. Early virologic failure is defined as two consecutive HIV-1 RNA measurements within the first 16 weeks that are above the baseline HIV-1 RNA level or more than a 1 $\log_{10}$ rise above the subject's nadir level. During the 84 week follow-up period, 35 and 17 subjects virologically failed the two arms, 1 and 8 subjects were lost to follow-up, and 29 and 39 subjects were administratively censored at 84 weeks, respectively. Table 2 displays the values of the test statistics and the corresponding p-values. The new test statistics $\widetilde{G}_{0.5}^2$, $\widetilde{T}^*$, $\widetilde{T}_0^*$, G_{max}, and $T(\mathcal{A})$ yield the smallest p-values, followed by the G^ρ test statistics, with the κ_1

and κ_0 tests giving the largest p-values. Thus, the new tests are more sensitive than the other tests for detecting a treatment effect. Given that the treatments appear to differ most in early and late periods (see Figure 1), these results are consistent with the simulation results.

To attempt to infer whether the observed treatment differences are due to early, late, or both early and late differences, we calculated the test statistics based on the linear combinations $G^1 - G^{-1}$, $\widetilde{G}^2_{0.5} - G^1$, $\widetilde{G}^2_{0.5} - G^{-1}$, $\widetilde{G}^2_{0.5} - G^0$, and $\widetilde{G}^2_{0.5} - (G^1 + G^{-1})/2$. The Z-statistics are 0.92, 1.68, 1.41, 2.08, and 1.59, respectively, with corresponding 2-sided p-values 0.36, 0.09, 0.16, 0.04, and 0.11. Thus, there is some evidence that the treatment differences are due to both early and late differences, and not merely to early differences or late differences.

We also performed two subgroup analyses, for those with baseline CD4 cell count less than its median and for those who previously used the drug Zidovudine (and thus may have drug resistant virus). The results are presented in Table 2. It is clear that, for these two subgroup analyses, the newly proposed tests $\widetilde{G}^2_{0.5}$, $\widetilde{T}^*, T(\mathcal{A}), G_{max}$, and $\widetilde{T}^*_0$ indicate significant treatment differences while the other tests (the G^ρ statistics and their supremum versions) do not. Notice that, in all analyses, the new statistic $\widetilde{G}^2_{0.5}$ gives the smallest p-values.

To conclude the example, we interpret the scientific meaning of the finding that the NFV + EFV regimen is superior to the EFV regimen in suppressing HIV replication, especially early and late in the follow-up period. Since early on virologic failure is due to the inability of therapy to initially suppress HIV replication, the NFV + EFV regimen is apparently more potent than the EFV regimen. This makes sense because the former regimen contains a protease inhibitor (NFV), whereas the latter regimen does not. Furthermore, the superiority of the NFV + EFV regimen late in follow-up may have been caused by a higher rate of development of drug resistance for subjects in the EFV arm. This interpretation is supported by findings that therapy with a protease-inhibitor-containing combined drug therapy delays the development of drug resistance (cf., Wong et al. 1997).

6 Discussion

To detect early and/or late survival differences which often occur in AIDS studies, we have developed new linear rank tests that offer appreciable power gains over the commonly used log-rank tests. When the investigator has strong reasons *a priori* to expect specific types of early and/or late survival differences, the newly proposed statistic $\widetilde{G}_a^2$, with suitable choices of the "turning index" a, is more sensitive than other statistics. With a suitable choice of a, $\widetilde{G}_a^2$ can be used to detect early differences, late differences, and early and late differences. When the investigator is somewhat uncertain about the specific types of alternatives, the newly proposed versatile test statistics $T(\mathcal{A})$ and $\widetilde{T}^*$ are more versatile than any of the individual statistics while being nearly as sensitive as the most powerful individual statistics. Moreover, the calculation of these statistics is relatively simple, compared to other previously proposed supremum statistics such as κ^0 and κ^1.

After the primary test of significance has been performed, the result must be interpreted. This interpretation can be guided by results from secondary analyses, such as those carried out with the linear combination statistics in Section 5. Such secondary analyses can be used to interpret any differences that may be observed among the results of various tests.

The powers of $T(\mathcal{A})$ and $\widetilde{T}^*$ may depend on the choice of $\mathcal{A}$ with regard to the richness of $\mathcal{A}$ and the inclusion of the true log hazard ratio. Golub (1963) suggested, however, that the weight functions actually used need not correspond exactly to the true log hazard ratio in order for most of the power gains to be realized. When we choose a finite set for $\mathcal{A}$, if the individual test statistics in this set are strongly correlated, then tests $T(\mathcal{A})$ and $\widetilde{T}^*$ based on a larger set may not add much power but may unnecessarily increase computational burden.

As pointed out by the Associate Editor, the parameters ρ and γ in the $G^{\rho,\gamma}$ family can be negative, in which case the $G^{\rho,\gamma}$ statistics can produce weight functions similar in shape to the quadratic weight function proposed in this paper. However, the proposed quadratic weight function $W_a(s)$ is simpler and easier to interpret. For example, as noted in Section 2, the value of a is closely related to the minimum weight and the ratio of the weights given to early and late events. Moreover, when the estimated survival function $\widehat{S}(s)$ attains values of 0 or 1, the $G^{\rho,\gamma}$ statistics with negative ρ and γ can become infinite, while this is not true for the proposed test statistics. The quadratic weight function is also similar in form to the weight function of Buyske et al. (2000). However, the proposed quadratic weight function is more flexible and is more appropriate for early and late survival differences, while the weight function in Buyske et al. (2000) is more similar to the usual G^{ρ} family except that it is designed for the case of rare events.

7 Appendix

As noted in Section 2, the proposed quadratic weight $W_a(t)$ is not predictable, so the standard martingale theory may not be directly applied to prove the asymptotic normality of the new test statistics. Following Buyske et al. (2000), here we outline a proof which shows that the random weight function $W_a(t)$ can be replaced by its nonrandom limit without altering the limit distributions of the test statistics. Let $P(t) = (1/\sqrt{n})\int_0^t Y_1(s)Y_2(s)/(Y_1(s) + Y_2(s))\{dN_1(s)/Y_1(s) - dN_2(s)/Y_2(s)\}$ and let $W_a^*(s) = S_1(s-) - (aS_1(\tau-) + 1 - a)$. Under $H_0 : S_1 = S_2$, we have $P(t) \to Z_P(t)$ weakly, where $Z_P(t) \sim N(0, V_P(t))$ and $V_P(t) = \int_0^t Y_1(s)Y_2(s)/(Y_1(s) + Y_2(s))^2\{dN_1(s) + dN_2(s)\}$, and $\sup_{0\le t\le\tau}|P(t) - Z_P(t)| \to 0$ a.s. and $\sup_{0\le t\le\tau}|\widehat{S}(t) - S_1(t)| \to 0$ a.s. Then, by integration by parts,

$$\begin{aligned}\frac{1}{n}\widetilde{G}_a^2 &= \int_0^\tau W_a(t)dP(s) = (1-a)(\widehat{S}(\tau-)-1)P(\tau) - \int_0^\tau P(s)dW_a(s)\\ &= (1-a)(S_1(\tau-)-1)Z_P(\tau) - \int_0^\tau Z_P(s)dW_a^*(s) + o(1) = \int_0^\tau W_a^{*2}(s)dZ_P(s) + o(1).\end{aligned}$$

Also, by the law of large number, we have $\widehat{\sigma}^2_{\widetilde{G}^2_a}/n \to \int_0^\tau W_a^{*4}(s)dV_P(s)$. Therefore, under H_0, we have $\widetilde{G}^2_a/\widehat{\sigma}_{\widetilde{G}^2_a} \to N(0,1)$ in distribution. The asymptotic result under the contiguous alternatives can be derived in exactly the same way.

ACKNOWLEDGMENTS

We are grateful to Victor DeGruttola, Bob Gray, and David Harrington for very helpful comments, and to the editor, associate editor, and two referees for very constructive suggestions. We thank the ACTG 364 Study Team for providing the data for the AIDS clinical trial example, including Mary Albrecht, Ronald Bosch, and David Katzenstein. This research was supported by a Canada NSERC grant and US grants 5-U01-AI38855 and 5-U01-AI28076 from the National Institutes of Health, NIAID.

REFERENCES

Buyske, S., Fagerstrom, R., and Ying, Z. (2000). A class of weighted log-rank tests for survival data when the event is rare. *J. Am. Statist. Assoc.* **95**, 249-258.

Davison, A.C. and Hinkley, D.V. (1997). *Bootstrap Methods and Their Application.* Cambridge University Press.

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

Fleming, T.R., Harrington, D.P., and O'Sullivan, M. (1987). Supremum versions of the log-rank and generalized Wilcoxon statistics. *J. Am. Statist. Assoc.* **82**, 312-320.

Gastwirth, J.L. (1966). On robust procedures. *J. Am. Statist. Assoc.* **61**, 929-948.

Genz, A. (1992). Numerical Computation of Multivariate Normal Probabilities. *Journal of Computational and Graphical Statistics* **1**, pp. 141-149.

Gilbert, P.B., Ribaudo, H.J., Greenberg, L., Yu, G., Bosch, R., and Tierney, C. (2000). Considerations in choosing a primary endpoint that measures durability of virologic suppression in an antiretroviral trial. *AIDS* **14**, 1-12.

Gilbert, P.B., Wei, L.J., Kosorok, M.R., and Clemens, J.D. (2002). Simultaneous inference on the contrast of two hazard functions. *Biometrics*, submitted.

Golub, G.H. (1963). Comparison of the variance of minimum variance and weighted least squares regression coefficients. *Annals of Mathematical Statistics* **43**, 984-991.

Gray, R.J. and Tsiatis, A.A. (1989). A linear rank test for use when the main interest is in differences in cure rates. *Biometrics* **45**, 899-904.

Guidelines for the use of antiretroviral agents in HIV-infected adults and adolescents (1998). *MMWR Morbidity and Mortality Weekly Report* **47**, 43-82.

Luenberger, D.G. (1984). *Linear and Nonlinear Programming*, 2nd ed. Reading Mass: Addison-Wesley.

Self, S.G. (1991). An adaptive weighted logrank test with application to cancer prevention and screening trials. *Biometrics* **47**, 975-986.

Shen, Y. and Cai, J. (2001). Maximum of the weighted Kaplan-Meier tests with application to cancer prevention and screening trials. *Biometrics* **57**, 837-843.

Stablein, D.M., Carter, W.H., and Novak, J.W. (1981). Analysis of survival data with nonproportional hazard functions. *Controlled Clinical Trials* **2**, 149-159.

Wong, J.K., Hezareh, M., Gunthard, H.F., Havlir, D.V., Ignacio, C.C., Splina, C.A., and Richman, D.D. (1997). Recovery of replication-competent HIV despite prolonged suppression of plasma viremia. *Science* **278**, 1291–1295.

Zucker, D.M. and Lakatos, E. (1990). Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. *Biometrika* **77**, 853-864.

List of figures and tables

Figure 1. Log hazard ratio estimate (solid line) with 95% pointwise and simultaneous confidence intervals for the EFV and EFV + NFV treatments in ACTG Study 364.

Figure 2. The (scaled) quadratic weight function $W_a(t)$, plotted against $F(t)$, for $a = 0, 0.25, 0.5, 0.75, 1$.

Figure 3. The four survival configurations in the alternative cases considered in the simulations. The λ_i are the parameters in the exponential distributions for treatment $i(i = 1, 2)$. The parameter values for the 20% (60%) censoring case are: for case I, $\lambda_1 = 1.2(0.4)$ and $\lambda_2 = 0.8(0.25)$; for case II, $\lambda_1 = 2.0(2.0), 0.4(0.16), 0.8(0.12)$ and $\lambda_2 = 0.4(0.16), 2.0(2.0), 0.8(0.12)$ for $t < 0.43(0.27), 0.43(0.27) \leq t < 0.86(0.54), t \geq 0.86(0.54)$ respectively; for case III, $\lambda_1 = 0.6(0.24), 2.5(2.0)$, and $\lambda_2 = 0.6(0.24), 1.0(0.3)$ for $t < 1.4(1.6), t \geq 1.4(1.6)$ respectively; for case IV, $\lambda_1 = 2.0(2.0), 0.1(0.2), 0.6(0.05), 2.0(0.8)$ and $\lambda_2 = 0.1(0.2), 2.0(2.0), 0.6(0.05), 0.8(0.08)$ for $t < 0.2(0.2), 0.2(0.2) \leq t < 0.4(0.4), 0.4(0.4) \leq t < 0.8(1.4), t \geq 0.8(1.4)$ respectively.

Table 1. Simulation Results for Sizes and Powers ($\times 100$) with $\alpha = 0.05$.

Table 2. Test Statistics and Two-sided p-values for the ACTG Study 364 Example.

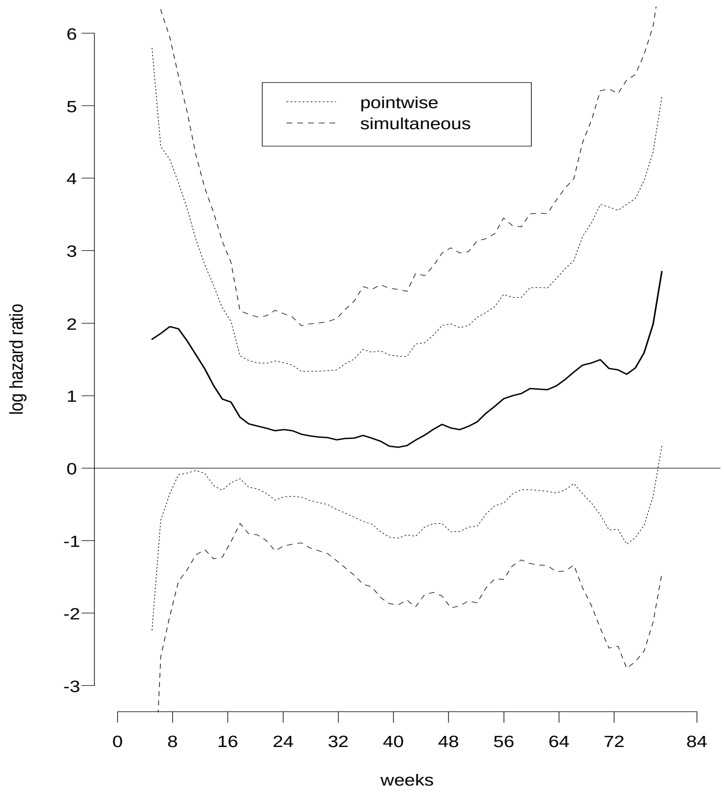

Figure 1:

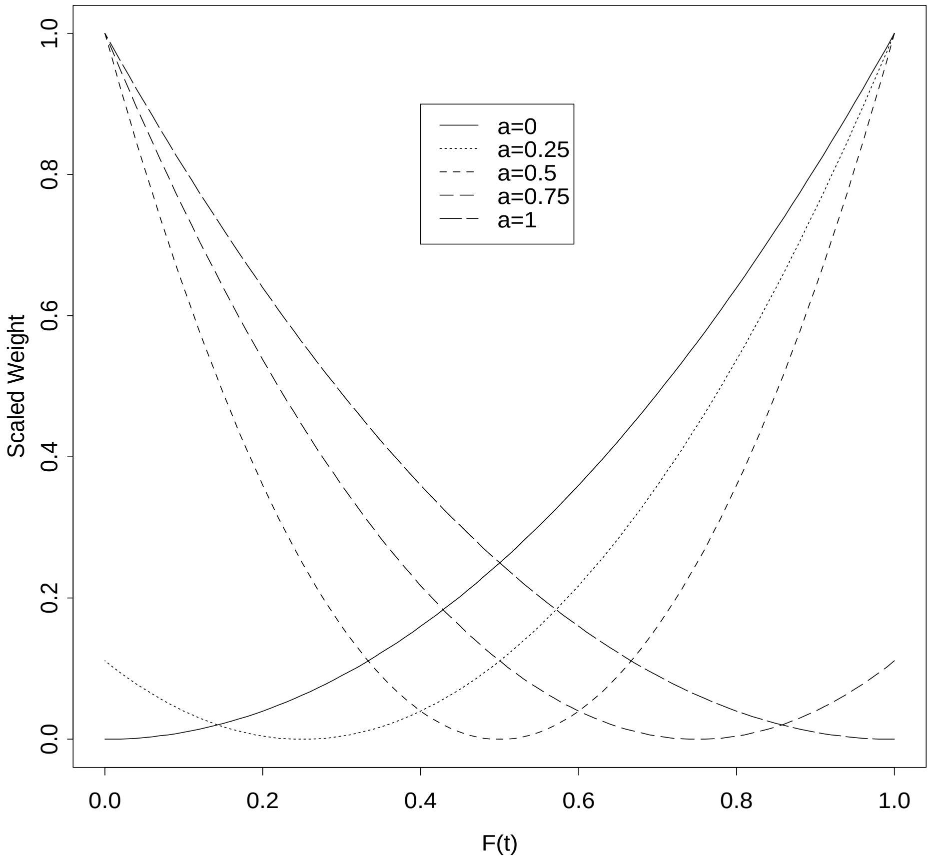

Figure 2:

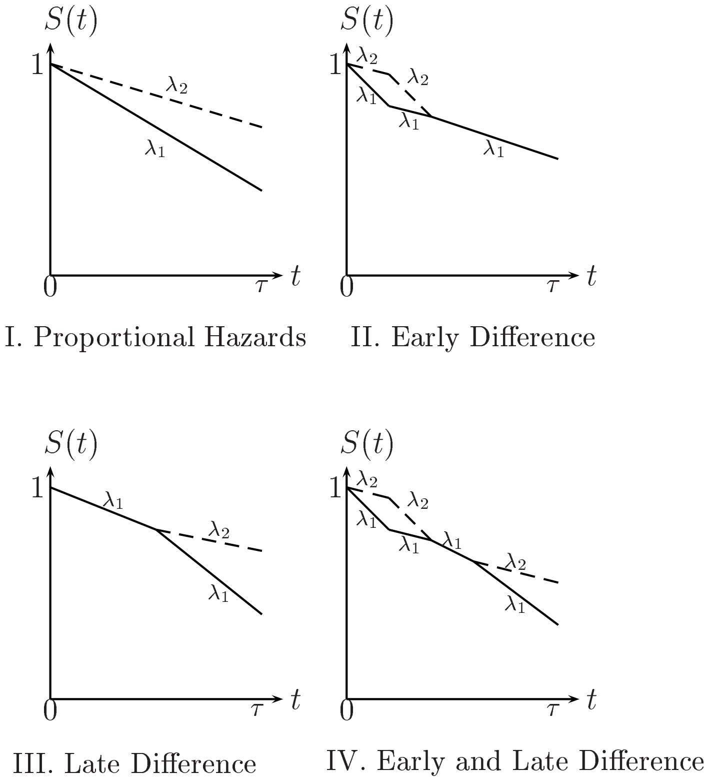
S(t)
1
λ2
λ1
0
τ
t
I. Proportional Hazards
II. Early Difference
III. Late Difference
IV. Early and Late Difference

Figure 3.

Table 1. Simulation Results for Sizes and Powers ($\times 100$) with $\alpha = 0.05$

Configuration	Censor.	$n_1 = n_2$	G^0	G^1	G^{-1}	$\widetilde{G}^2_1$	$\widetilde{G}^2_0$	$\widetilde{G}^2_{0.5}$	$\widetilde{G}^2_{0.4}$	κ_0	κ_1	$T(\mathcal{A})$	G_{max}	$\widetilde{T}^*_0$	$\widetilde{T}^*$
NULL	20%	20	5.8	5.1	6.2	4.9	6.7	5.4	6.5	4.7	4.6	6.0	6.0	5.1	7.4
		50	5.5	5.3	4.2	6.0	4.4	5.2	4.4	5.1	5.5	5.0	5.4	5.4	6.6
	60%	20	4.4	4.1	5.3	4.4	5.2	4.4	4.8	3.2	3.1	4.8	5.1	5.1	6.4
		50	5.9	5.4	6.1	5.9	4.3	6.6	6.0	4.1	4.8	5.8	6.1	6.1	7.0
I (prop.hazard)	20%	20	30	26	26	20	22	22	22	26	22	24	28	30	28
		50	56	51	49	38	41	40	38	53	45	42	53	55	48
	60%	20	24	24	24	18	20	18	17	20	20	19	24	24	22
		50	44	43	43	31	27	32	27	39	37	32	41	42	38
II (early diff.)	20%	20	19	42	7	63	2	33	8	40	65	58	35	24	47
		50	32	75	7	96	1	70	12	93	98	96	75	45	85
	60%	20	11	18	5	60	0	14	2	23	39	51	15	11	36
		50	14	31	6	91	0	42	4	84	92	93	40	22	72
III (late diff.)	20%	20	19	10	33	5	37	24	37	16	9	31	26	22	31
		50	35	16	63	7	68	45	66	29	13	59	53	40	58
	60%	20	35	28	42	8	59	36	53	28	22	47	40	39	47
		50	60	46	72	8	91	62	87	52	39	84	68	64	83
IV (early & late)	20%	20	25	23	33	29	30	68	56	20	30	60	57	40	60
		50	48	44	61	55	53	98	92	52	84	94	94	74	95
	60%	20	30	31	29	38	27	73	55	23	26	70	64	44	69
		50	56	58	55	72	51	99	92	62	72	98	98	82	98
Average Power	20%	20	23	25	25	29	23	37	31	26	32	43	36	29	42
		50	43	46	45	49	41	63	52	57	60	73	69	54	72
	60%	20	25	25	25	31	26	35	32	24	27	47	36	30	44
		50	44	44	44	50	42	59	52	59	60	77	62	52	73

Table 2. Test Statistics and Two-sided p-values for the ACTG Study 364 Example

	$\widetilde{G}^2_{0.5}$	$\widetilde{T}^*$	$\widetilde{T}^*_0$	G_{max}	$T(\mathcal{A})$	G^1	G^0	G^{-1}	κ_1	κ_0
Test statistic	3.57	3.56	3.03	3.57	3.57	3.16	3.11	3.03	3.15	13.11
p-value ($\times 10^4$)	3.6	4.5	5.6	6.1	12.5	15.8	18.5	24.1	32.8	37.1
Ratio of p-values*	1.0	1.2	1.6	1.7	3.4	4.4	5.1	6.6	9.0	10.0
	Subgroup Analysis (Baseline CD4 Cell Count < median, $n_1 = 34, n_2 = 30$)									
	$\widetilde{G}^2_{0.5}$	$\widetilde{T}^*$	$T(\mathcal{A})$	$\widetilde{T}^*_0$	G_{max}	G^1	G^0	κ_1	κ_0	G^{-1}
Test statistic	2.41	2.64	2.68	2.20	2.40	2.07	1.92	2.15	2.07	1.76
p-value	0.016	0.016	0.022	0.028	0.032	0.038	0.054	0.062	0.077	0.078
	Subgroup Analysis (Prior Use of Zidovudine, $n_1 = 33, n_2 = 35$)									
	$\widetilde{G}^2_{0.5}$	$\widetilde{T}^*$	$T(\mathcal{A})$	G_{max}	$\widetilde{T}^*_0$	G^{-1}	G^0	G^1	κ_0	κ_1
Test statistic	2.93	3.10	3.15	2.93	2.58	2.44	2.20	1.96	2.21	1.98
p-value	0.003	0.004	0.005	0.007	0.010	0.015	0.027	0.050	0.053	0.094

*Relative to the p-value for the test based on $\widetilde{G}^2_{0.5}$.