Published in final edited form as:
Biometrics. 2010 March ; 66(1): 30–38. doi:10.1111/j.1541-0420.2009.01243.x.

Improved Logrank-Type Tests for Survival Data Using Adaptive Weights

Song Yang and
Office of Biostatistics Research, National Heart, Lung, and Blood Institute, 6701 Rockledge Dr. MSC 7913, Bethesda, Maryland 20892, U. S. A. yangso@nhlbi.nih.gov

Ross Prentice
Fred Hutchinson Cancer Research Center, 1100 Fairview Ave. N., M3-A410, Seattle, WA 98109, U. S. A. rprentic@whi.org

SUMMARY

For testing for treatment effects with time to event data, the logrank test is the most popular choice and has some optimality properties under proportional hazards alternatives. It may also be combined with other tests when a range of nonproportional alternatives is entertained. We introduce some versatile tests that use adaptively weighted logrank statistics. The adaptive weights utilize the hazard ratio obtained by fitting the model of Yang and Prentice (2005). Extensive numerical studies have been performed under proportional and nonproportional alternatives, with a wide range of hazard ratios patterns. These studies show that these new tests typically improve the tests they are designed to modify. In particular, the adaptively weighted logrank test maintains optimality at the proportional alternatives, while improving the power over a wide range of nonproportional alternatives. The new tests are illustrated in several real data examples.

Keywords

Clinical trials; Proportional and non-proportional hazards; Survival analysis; Time-varying treatment effect; Weighted logrank tests

1. Introduction

The logrank test has been the method of choice for testing for existence of a treatment effect with survival data (Mantel, 1966; Peto and Peto, 1972). It is (asymptotically) optimal under proportional hazards alternatives, with equal censoring patterns in the two groups. In this work, we show that the logrank test and related tests can be improved by using weighted logrank statistics with adaptive weights.

The key requirement for the optimality of the logrank test is the proportional hazards assumption. This assumption provides a suitable approximation in many situations, and the hazard ratio estimates from a proportional assumption can provide simple and useful summary measures, even if the hazard ratio is moderately time-dependent (e. g. Prentice, Pettinger, and Anderson, 2005). Correspondingly, in those situations the logrank test typically has good power.

SUPPLEMENTARY MATERIALS
The Appendix referenced in Section 2.3 is available under the Paper Information link at the *Biometrics* website http://www.biometrics.tibs.org.

There are a variety of situations when the nonproportionality of the hazard functions is severe. The treatment may bring a short-term benefit, and gradually lose its effect as time goes on. This results in converging hazard functions that often can be fit well by the proportional odds model, a well-known alternative to the proportional hazards model (Bennett, 1983). Conversely, it may take a relatively long time for a "conservative" treatment to fully realize its effect. Such a lag in the initiation of a treatment effect could result in little or no difference in early survival experiences in the groups being compared, but an increasingly noticeable difference or even divergence during the later part of the follow-up period. An "aggressive" treatment, on the other hand, may result in higher mortality early on due to toxicity or complications, but may provide a beneficial effect in the long run. It may even be possible for a treatment to be beneficial initially and then turn harmful in the long run. In general, the longer the follow-up period is, the more likely it is for various nonproportional scenarios to develop.

If a nonproportional alternative can be pre-specified, then a weighted logrank test can be used, with the weight chosen appropriately to maximize the power. For example, a priori estimate of the weight can be used based on the clinician's pretrial projections to improve power over the logrank test (Lagakos, Lim, and Robins, 1990; Zucker, 1992). The Peto-Prentice test (Peto and Peto, 1972; Prentice, 1978), and in general the $G^{\rho,\gamma}$ test (Fleming and Harrington, 1991), can also be recast as weighted logrank tests. Under mild conditions, weighted logrank tests are consistent under ordered hazards alternatives and, if the weight is monotone, under stochastic ordering alternatives (Gill, 1980, Ch. 4).

In applications, it may often be desirable to design a test with good power over a range of possible alternative hypotheses. For such a purpose, combinations of weighted logrank tests may be considered. Examples include the linear combinations (Gastwirth, 1985; Zucker and Lakatos, 1990), maximum of several standardized tests (Tarone, 1981; Fleming and Harrington, 1984, 1991), maximum after orthogonalization (Breslow et al., 1984), or χ^2 test of the simultaneous null hypothesis for the set of tests. For the maximum test approach, in addition to taking the maximum over a finite number of tests, the maximum can be taken over an infinite collection of tests (Self, 1991) or over time (Gill, 1980; Fleming and Harrington, 1984; Fleming, Harrington, and O'Sullivan, 1987), or in general over a class of function-indexed tests (Kosorok and Lin, 1999). In those more complex situations, often Monte Carlo methods are used to obtain the theoretically intractable critical value and p-value. Typically the combination approaches are more robust than the logrank test, in the sense of maintaining good power under a range of nonproportional hazards alternatives. Usually the combination tests fail to remain optimal under proportional hazards alternatives, giving one more example of the trade-off between efficiency and robustness seen in many statistical problems. For the transformed two sample location model, an asymptotically efficient test can be obtained (Lai and Ying, 1990). It uses kernel estimates and requires large sample sizes to perform well. Pecova and Fleming (2003) proposed a maximum test using as the selector an estimator of the asymptotic relative efficiency for a finite collection of baseline distributions in the location model. The test is asymptotically efficient among the finite collection it is based on. But simulation studies show that, for small samples, there is some loss of power for the proportional hazards alternatives compared with the logrank test.

In this work, we propose to modify the logrank and related tests by using adaptive weights. Under proportional hazards alternatives, the new adaptively weighted logrank test continues to be optimal. When the hazards are nonproportional, the adaptive weights reflect deviations from proportionality, and lead typically to improvement in power over the logrank test. To control the potentially inflated test size, an adjustment is proposed which takes into consideration the correlation between the relevant statistics. The adaptive weights are also used to modify the closely related maximum test and the Breslow test. The adaptive weights

are obtained by fitting the data to the model of Yang and Prentice (2005), which contains the proportional hazards model and the proportional odds model as submodels, and accommodates nonproportional hazards situations to the extreme of having crossing hazards and crossing survivor functions. Extensive simulation studies show that, for a wide variety of nonproportional hazards alternatives, the proposed modifications typically improve the power. Overall, the adjusted adaptively weighted logrank test has the best performance, maintaining optimality under the proportional alternatives, while improving the power under a wide range of nonproportional alternatives. We illustrate the newly proposed tests in diverse examples, ranging from mildly time-dependent hazard ratios, to more severe nonproportionality of the hazard functions, and to the extreme case of crossing survival functions. These motivating examples indicate that, in one degree or another, various robust alternatives to the logrank test may be more sensitive to substantial deviations from the proportional hazards condition, but may also be less sensitive when the nonproportionality is mild. The adjusted adaptively weighted logrank test, on the other hand, has a more stable behavior throughout all the examples.

We organize the paper as follows. In Section 2, after the notation and setup, we describe the model of Yang and Prentice (2005), which is used for obtaining the adaptive weights. Then the adaptively weighted logrank test is introduced. In Section 3, the adaptive weights are used to modify a few related tests in the literature. In Section 4, the new tests are examined in simulation studies for a wide range of hazard ratio patterns. In Section 5, the new tests are illustrated in several examples with diverse nonproportional behavior. A discussion is given in Section 6.

2. Adaptively Weighted Logrank Test

2.1 Preliminaries

Label the two groups control and treatment, with survivor functions S_C, S_T respectively. We consider the hypothesis testing problem

$$H_0: S_T = S_C \qquad vs. \qquad H_a: S_T \neq S_C. \tag{1}$$

Let $T_1, \ldots, T_n$ be the pooled lifetimes of the two groups, beginning with the control group, $Z_i = I(i > n_1)$, $i = 1, \cdots, n$, where $n_1 < n$ is the size of the control group and $I(\cdot)$ is the indicator function. Also, let $C_1, \ldots, C_n$ be the censoring variables. The available data consist of the triplets (X_i, δ_i, Z_i), $i = 1, \ldots, n$, where $X_i = \min(T_i, C_i)$ and $\delta_i = I(T_i \leq C_i)$. We assume that $T_1, \ldots, T_n, C_1, \ldots, C_n$ are independent, with $T_1, \ldots, T_{n_1}$, $n_1 < n$, having the survivor function S_C and $T_{n_1+1}, \ldots, T_n$, having the survivor function S_T. The censoring variables (C_i's) need not be identically distributed, and in particular the two treatment groups may have different censoring patterns.

Let

$$N_i(t) = \delta_i I(X_i \leq t), \; Y_i(t) = I(X_i \geq t), \tag{2}$$

and define

$$K(t) = \sum_{i=1}^{n} Y_i(t), \quad K_Z(t) = \sum_{i=1}^{n} Z_i Y_i(t). \tag{3}$$

Then, the weighted logrank statistic for the testing problem (1) is

$$U_\Psi = \sum_{i=1}^{n} \int_0^\infty \Psi(t) \left\{ Z_i - \bar{Z}(t) \right\} dN_i(t), \tag{4}$$

where $\bar{Z}(t) = K_Z(t)/K(t)$, and Ψ is a possibly data-dependent nonnegative weight function satisfying certain regularity conditions. We will denote the corresponding standardized test statistic by W_Ψ. Hence $W_\Psi = U_\Psi / \sqrt{V_\Psi}$, where $V_\Psi = \sum_{i=1}^{n} \int_0^\infty \Psi^2(t) [\{K(t) - K_Z(t)\} K_Z(t)/K^2(t)] dN_i(t)$. In particular, the standardized logrank is denoted by W_1.

Assume that the distributions of the failure times are absolutely continuous and λ_C, λ_T are the hazard functions for the two groups respectively. Suppose λ_C, λ_T belong to a parametric family $\{\lambda_\theta, \theta \in \Theta\}$ and $\lambda_C = \lambda_{\theta_0}$. For a sequence of local alternatives $H_a^n: \lambda_T = \lambda_{\theta_n}$, it is well known that (cf. Gill, 1980; Fleming and Harrington, 1991), under appropriate regularity conditions, the weight that maximizes the power under H_a^n has a limit that is proportional to

$$\frac{\partial}{\partial \theta} \log \lambda_\theta |_{\theta=\theta_0}. \tag{5}$$

Furthermore, this weight results in full efficiency when $\pi_1 = \pi_2$, where π_1, π_2 are the limits of $(K - K_Z)/n_1$, K_Z/n_2 respectively. Thus the logrank statistic, corresponding to $\Psi \equiv 1$, is optimal when $\{\lambda_\theta, \theta \in \Theta\}$ follows the proportional hazards model: $\lambda_\theta = \theta \lambda_0$.

2.2 The Model of Yang and Prentice (2005)

To improve the logrank test for a range of alternative hypotheses, it seems natural to consider a model that extends the proportional hazards model and accommodates a variety of nonproportional hazards situations. Let

$$\tau_0 = \sup \{t: S_C(t) > 0\}. \tag{6}$$

Recently Yang and Prentice (2005) proposed a model in which

$$\lambda_T(t) = \frac{\theta_1 \theta_2}{\theta_1 + (\theta_2 - \theta_1) S_C(t)} \lambda_C(t), \ t < \tau_0, \tag{7}$$

where θ_1 and θ_2 are positive. This model contains the proportional hazards model ($\theta_1 = \theta_2$) and the proportional odds model ($\theta_2 = 1$). Under this model, $\theta_1 = \lim_{t \downarrow 0} \lambda_T(t)/\lambda_C(t)$, $\theta_2 = \lim_{t \uparrow \tau_0} \lambda_T(t)/\lambda_C(t)$. Thus θ_1 and θ_2 can be interpreted as the short-term and long-term hazard ratios, respectively. Various combinations of θ_1 and θ_2 give different nonproportional hazards patterns, such as disappearing treatment effect ($\theta_2 = 1$), or no initial effect ($\theta_1 = 1$). When $\theta_1 \neq \theta_2$ and the interval formed by θ_1 and θ_2 contains one, the two hazard functions cross. The survivor functions S_C and S_T also cross if either $\theta_1 < 1$ and $\theta_2 > 1$, or $\theta_1 > 1$ and $\theta_2 < 1$. In comparison, the survival functions will not cross under the proportional hazards models, the proportional odds model and, in general, under linear transformation models (Bickel et al. 1993). Neither will they cross under the accelerated failure time model. Thus this model allows more flexible patterns for the treatment effect compared with some of the traditional models.

To test the hypothesis of no treatment effect, Yang and Prentice (2005) proposed a χ^2 test using their two estimating functions. That test turns out to be the χ^2 test that combines the logrank statistic and Peto-Prentice statistic. Alternatively, one can define $\beta_1 = \log \theta_1$ and $\beta_2 = \log \theta_2$, and set $\beta_1 = \gamma_1\theta$ and $\beta_2 = \gamma_2\theta$ and consider (5) for $\theta_0 = 0$. The resulting test is of the form (4) with weight function

$$\tilde{\Phi}=\gamma_1 S_C+\gamma_2(1-S_C). \tag{8}$$

For the proportional hazards submodel with $\gamma_1 = \gamma_2$, $\tilde{\Phi}$ reduces to a constant, thus the logrank test is optimal; while for the proportional odds model with $\gamma_2 = 0$, Φ becomes $\gamma_1 S_C$, thus the weight S_C is optimal, confirming well known results in the literature.

2.3 Adaptively Weighted Logrank Test

Since $\tilde{\Phi} = \lim(1 - \lambda_C/\lambda_T)/\theta$ as $\theta \downarrow 0$, we can consider a test with the simple weight function $\Phi_1 = \hat{\lambda}_C/\hat{\lambda}_T$, where the estimated hazard functions are obtained by fitting the model of Yang and Prentice (2005) to the data. This test reduces to the logrank test under the proportional hazards model $\beta_1 = \beta_2$. For symmetry in the two treatments, we can derive an adaptive test that also uses $\Phi_2 = 1/\Phi_1$. Note that, to use these weights, we need to restrict to the case where the hazard ratio is not zero. Under nonproportional alternatives, it is plausible that either Φ_2 or Φ_1 would often be more sensitive to departure from the null hypothesis than a constant weight, and consequently could result in improved power.

These weight functions depend on the estimated hazard ratio under model of Yang and Prentice (2005). We refer the readers to that paper for the motivation of the estimated hazard ratio there. Here we give the technical details necessary to obtain the weight functions. Let

$$H_j(t;\mathbf{b})=\sum_{i=1}^{n}\delta_i\gamma_{ji}(\mathbf{b})I(X_i \le t),\ j=1,2, \tag{9}$$

for $t > 0$, where $\gamma_{ji}(\mathbf{b}) = \exp(-b_j Z_i)$ and $\mathbf{b} = (b_1, b_2)$. Using these functions and $K(t)$ in (2), define $\widehat{\Lambda}_1(t;\mathbf{b})=\int_0^t \frac{1}{K(s)}\widehat{H}_1(ds;\mathbf{b})$, $\widehat{\Lambda}_2(t;\mathbf{b})=\int_0^t \frac{1}{K(s)}\widehat{H}_2(ds;\mathbf{b})$, and

$$\widehat{R}(t;\mathbf{b})=\frac{1}{\widehat{P}(t;\mathbf{b})}\int_0^t \widehat{P}_-(s;\mathbf{b})\widehat{\Lambda}_1(ds;\mathbf{b}), \tag{10}$$

where $\hat{P}(t;\mathbf{b}) = \Pi_{s\le t}\{1-\Delta\hat{\Lambda}_2(s;\mathbf{b})\}$, with $\hat{P}_-$ denoting the left continuous, in t, version of $\hat{P}$, and $\Delta\hat{\Lambda}_2(s;\mathbf{b})$ the jump of $\hat{\Lambda}_2$ in s.

Now define

$$\widehat{M}_i(t;\mathbf{b})=\delta_i I(X_i \le t)-\int_0^t I(X_i \ge s)\frac{\widehat{R}(ds;\mathbf{b})}{\gamma_{1i}(\mathbf{b})+\gamma_{2i}(\mathbf{b})\widehat{R}(s;\mathbf{b})},\ 1 \le i \le n. \tag{11}$$

The estimating function $Q(\mathbf{b})$ proposed in Yang and Prentice (2005) is, for some $\tau < \tau_0$,

$$Q(\mathbf{b})=\sum_{i=1}^{n}\int_0^\tau f_i(t;\mathbf{b})\widehat{M}_i(dt;\mathbf{b}), \tag{12}$$

with

$$f_{1i}(t;\mathbf{b})=Z_i\frac{\gamma_{1i}(\mathbf{b})}{\gamma_{1i}(\mathbf{b})+\gamma_{2i}(\mathbf{b})\widehat{R}(t;\mathbf{b})},\quad f_{2i}(t;\mathbf{b})=Z_i\frac{\gamma_{2i}(\mathbf{b})\widehat{R}(t;\mathbf{b})}{\gamma_{1i}(\mathbf{b})+\gamma_{2i}(\mathbf{b})\widehat{R}(t;\mathbf{b})}. \tag{13}$$

Let $\hat{\beta} = (\hat{\beta}_1, \hat{\beta}_2)$ be the zero of $Q(\mathbf{b})$. Then $\hat{\beta}$ is the pseudo maximum likelihood estimator of (β_1, β_2). Thus the adaptive weight Φ_2, or the estimated hazard ratio under the model of Yang and Prentice (2005), is

$$\Phi_2(t)=\frac{1+\widehat{R}(t;\widehat{\beta})}{\exp(-\widehat{\beta}_1)+\exp(-\widehat{\beta}_2)\widehat{R}(t;\widehat{\beta})}. \tag{14}$$

From this, Φ_1 follows easily.

Using the standardized statistics W_{Φ_1}, W_{Φ_2} corresponding to the weights Φ_1, Φ_2, we can define a test that rejects H_0 when

$$\max(|W_{\Phi_1}|, |W_{\Phi_2}|)>z(\alpha/2), \tag{15}$$

where $z(\alpha/2)$ is the upper $100(1-\alpha/2)$th percentile of the standard normal distribution.

Note that, due to the dependence on $\hat{\beta}$, the weights Φ_1, Φ_2 do not satisfy the typical requirement for the usual weighted logrank tests. In the Appendix, given in the Supplementary Materials, we show that, under certain regularity conditions, W_{Φ_i}, $i = 1, 2$ are asymptotically equivalent to each other and to the standardized logrank statistic, under both H_0 and the proportional hazards alternatives. Thus not only does the test (15) have asymptotically correct size, but it is also asymptotically optimal for the proportional hazards alternatives. For nonproportional alternatives, it is likely that either Φ_1 or Φ_2 would behave more closely to $\tilde{\Phi}$ than a constant weight, and hence is expected to result in better power than the logrank test. To control the size, a multiple comparison adjustment to the test (15) will be described in the next section.

For one sided alternative hypotheses, the adaptive test can be modified correspondingly. Consider the alternative $H_a^{'}:S_T \geq S_C$, where the strict inequality holds for an interval over the real line. Then a test analogous to (15) is to reject H_0 when

$$\max(-W_{\Phi_1}, -W_{\Phi_2})>z(\alpha).$$

For the alternative $H_a^{''}:S_T \leq S_C$, where the strict inequality holds for an interval over the real line, the test analogous to (15) is to reject H_0 when

$$\max(W_{\Phi_1}, W_{\Phi_2})>z(\alpha).$$

3. Other Related Tests

3.1 The Maximum Tests

The adaptively weighted logrank statistics U_{Φ_1} and U_{Φ_2} can also be used in combination with other statistics to obtain new tests. With the logrank test and an additional weighted logrank statistic W_Ψ, the usual maximum test rejects H_0 when

$$\max(|W_1|, |W_\Psi|) > c_\rho(\alpha), \tag{16}$$

where the critical value c_ρ depends on the correlation coefficient ρ between W_1 and W_Ψ and can be computed using the asymptotic bivariate normal distribution of (W_1, W_Ψ). Note that for two weighted logrank statistics, the correlation coefficient ρ is nonnegative. For the level $\alpha = 0.05$, Table 1 in Yang, Hsu, and Zhao (2005) gives the value of $c_{|\rho|}(\alpha)$ for a range of correlation coefficient between two asymptotically normal tests, so that for a given ρ, the value of $c_{|\rho|}$ can be obtained through interpolation. Alternatively, the exact values can be obtained using bivariate normal calculations as given in (2) of Yang et al. (2005).

Let ρ_i be the correlation coefficient between W_{Φ_i} and W_Ψ, $i = 1, 2$. For the maximum test in (16), one obvious modification is to reject H_0 if

$$\max(|W_{\Phi_1}|, |W_{\Phi_2}|, |W_\Psi|) > c_\rho(\alpha), \tag{17}$$

where we choose $\rho = \min(\rho_1, \rho_2)$, in order to correct for the potentially inflated size. Alternatively, asymptotically exact critical values can be obtained from trivariate normal calculations. Numerical results show that the simpler approach in (17) is adequate. For combining more than two tests, the modification is analogous.

The critical value $c_\rho(\alpha)$ can also be used to provide an adjustment to (15) to improve the asymptotic distributional approximation: Let $\tilde{\rho}$ be the correlation coefficient between W_{Φ_1} and W_{Φ_2}. Then an adjusted version of (15) is to reject H_0 when

$$\max(|W_{\Phi_1}|, |W_{\Phi_2}|) > c_{\tilde{\rho}}(\alpha). \tag{18}$$

For a given data set, the significance level P for the tests in (16) through (18) can be obtained by

$$1 - P = \int_{-t}^{t} \left\{ F\left(\frac{t - \rho w}{\sqrt{1-\rho^2}}\right) - F\left(\frac{-t - \rho w}{\sqrt{1-\rho^2}}\right) \right\} f(w)dw, \tag{19}$$

where t is the observed test statistic value, ρ the estimated correlation coefficient, and f, F the density and cumulative distribution functions of the standard normal distribution respectively.

3.2 The Modified Breslow Test

As a complement to the logrank test, Brelsow et al. (1984) introduced a score test for the null hypothesis of proportional hazards against rank-regression alternatives. The test was motivated by situations where the hazard functions cross over, giving a treatment effect that Brelsow et al. (1984) described as an acceleration of the events of interest. Their test is related to the two sample model

$$\lambda_T = \exp\{\alpha + \beta z(t)\}\lambda_C, \tag{20}$$

where $z(t)$ is a time-dependent covariate. This model was proposed by Cox (1972) to cope with more complicated relationships than proportional hazards, and to check the two sample proportional hazards assumption. The test of Brelsow et al. (1984) is the score test of the composite null hypothesis $\beta = 0$ versus alternatives $\beta \neq 0$ for model (20). Let A_α, A_β be the scores for model (20), and define $\hat{\alpha}$ to be the partial likelihood estimator of α when $\beta = 0$. Then the test statistic for acceleration is $A_\beta(\hat{\alpha}, 0)$. Let X_z be the standardized test statistic. For testing H_0, the test statistic of Breslow et al. (1984) is $\max(|W_1|, |X_z|)$. Since W_1 and X_z are asymptotically independent under H_0, the critical value of this test can be easily obtained.

Breslow et al. (1984) investigated in detail the two special cases with the rank scores $z(T_i) = \#\{j : T_j \leq T_i, \delta_j = 1\}$ and the cumulative-hazard scores $z(T_i) = \hat{\Lambda}(T_i)$, where $\hat{\Lambda}$ is the Nelson-Aalen cumulative hazard estimator under H_0. Simulations and real data examples show that their test provides improved power against acceleration alternatives (i.e. with crossing hazards), compared with the logrank test and the Peto-Prentice test.

Recall N_i, Y_i, $i = 1, \cdots, n$ from (2) and define

$$K(t, \alpha) = \sum_{i=1}^{n} \exp(\alpha Z_i) Y_i(t), \; K_z(t, \alpha) = \sum_{i=1}^{n} Z_i \exp(\alpha Z_i) Y_i(t). \tag{21}$$

Let $U_\Psi(\alpha) = \sum_{i=1}^{n} \int_0^\infty \Psi(t)\{Z_i - \bar{Z}(t, \alpha)\} dN_i(t)$, where $\bar{Z}(t, \alpha) = K_Z(t, \alpha)/K(t, \alpha)$. The test statistic for acceleration can then be written as $U_\Psi(\hat{\alpha})$. Let $W_\Psi(\hat{\alpha})$ be the standardized version of $U_\Psi(\hat{\alpha})$. The test of Breslow et al. (1984) rejects H_0 if

$$\max\{|W_1|, |W_\Psi(\widehat{\alpha})|\} > z\left(\frac{1 - \sqrt{1-\alpha}}{2}\right), \tag{22}$$

where the critical value follows from the asymptotic independence of W_1 and $W_\Psi(\hat{\alpha})$. Similar to (15), a modification of this test is to reject H_0 if $\max(|W_{\Phi_1}|, |W_{\Phi_2}|, |W_\Psi(\widehat{\alpha}_1)|) > z\left\{(1 - \sqrt{1-\alpha})/2\right\}$, where $\hat{\alpha}_1$ is the zero of $U_{\Phi_1}(\alpha)$. However, as the maximum is taken over three individual tests, simulation results show that this test often has inflated size for moderate samples. An adjustment is to reject H_0 unless

$$|W_\Psi(\widehat{\alpha}_1)| \leq z\left(\tfrac{1-\sqrt{1-\alpha}}{2}\right) \quad \text{and} \quad \max(|W_{\Phi_1}|, |W_{\Phi_2}|) \leq c_{|\tilde{\rho}|}(1 - \sqrt{1-\alpha}). \tag{23}$$

For $\alpha = .05$, the critical value $c_{\tilde{\rho}}(1 - \sqrt{1-\alpha})$ is given in Table 1 for select values of $\tilde{\rho}$. For other values of ρ, one can either interpolate from this table, or use the exact method as in (2) of Yang et al. (2005) as mentioned before.

In the Appendix, we show that, under certain regularity conditions, asymptotically all of the newly proposed tests have the correct size. It is possible to have other pairs of weight functions that converge to constant functions under H_0. For example, instead of Φ_1, Φ_2, one could consider the ratios of Nelson-Aalen cumulative hazard functions, Kaplan-Meier cumulative distribution functions, or Kaplan-Meier survivor functions. The major attraction

of these alternative weight functions is the ease of computation. They converge to the constant one under H_0, and the corresponding tests also have the correct size asymptotically. Under non-proportional hazards alternatives, these weights likely would result in better power compared with the logrank test. However, since they are ratios of estimators from each single sample, for moderate sample sizes they may be unstable, as will be shown in the numerical studies described below.

4. Simulation Studies

We have performed extensive numerical studies to systematically examine the behavior of the new tests under a wide range of alternatives. Denote the tests in (15) through (18), (22) and (23) by LRAD, MX, MXAD, LRAD2, MXB, MXBAD respectively, with Ψ in (16), (17), (22) and (23) defined below. For the purpose of comparison, additional tests were also included in the numerical studies. Among them are the logrank test and the χ^2 test that combines the logrank test and the weighted logrank test with weight Ψ. The tests similar to (15) and (17), but with the ratios of estimated cumulative hazard functions instead of Φ_1, Φ_2, are denoted by CH, CH2 respectively. The ratios are defined to be zero where the denominator vanishes. The performance of the tests was examined for various combinations of short-term and long-term effects of the treatment, as follows:

(N). No treatment effect;

(PR+). Constant beneficial effect;

(PR−). Constant adverse effect;

(S0L+). No initial effect but a gradually increasing beneficial effect;

(S0L−). No initial effect but a gradually increasing adverse effect;

(S+L0). An initial beneficial effect that diminishes long-term;

(S−L0). An initial adverse effect that diminishes long-term;

(S−L+). Initial adverse effect but a long-term beneficial effect;

(S+L−). Initial beneficial effect but a long-term adverse effect;

(U). *U*- shaped treatment effect.

While it is impossible to exhaust all possible treatment effect scenarios, the above situations do include a reasonably wide range of possibilities considered in the literature. All except the last case correspond to various scenarios where the hazard ratio stays constant, gradually deviates from one, converges to one, or gradually increase or decrease to cross the line of constant one. Note that, under the first six alternatives (PR+) through (S−L0), the two distributions are stochastically ordered. Under the last three alternatives, without a stochastic ordering of the underlying distributions, it is not clear how to judge overall whether the treatment is better. These situations may be less common, but it is still important to have good power for testing the difference between the two distributions.

Now we report the results from a representative study. For cases (N) through (S+L−), the data were generated from the model of Yang and Prentice (2005), with $S_C(t)$ being the standard log-logistic, and with β being the zero vector, a multiple of (1, 1), (1, 0) or (0, 1), or having opposite signs in the two components respectively. For the case (U), the control group had the standard exponential distribution, and

$$\frac{\lambda_T(t)}{\lambda_C(t)} = \frac{1+a}{a}, \quad t \in (0, .5) \cup (1.5, \infty)$$
$$= \frac{a}{1+a}, \quad t \in [0.5, 1.5],$$

for some constant $a > 0$. The parameters in all the configurations were chosen so that the logrank test had approximately 70% power when $n_1 = n_2 = 80$ with 10% censoring. For all cases the censoring variables had a log- normal distribution, where the normal distribution had mean c and standard deviation 0.5, with c chosen to achieve various censoring rates. The weight Ψ in (16), (17), (22) and (23) was chosen, after examining several candidates, to be the Nelson-Aalen cumulative hazard estimator under H_0. For more stable tail behavior, Ψ was stopped at the order statistic near the 95th percentile. For various sample sizes and censoring levels, the simulation results based on 1000 repetitions are summarized in Table 2. The numbers given are the empirical size for case (N), and the power ratio over the log rank test for cases (PR+) through (U).

From these results, first we see that the CH test and the CH2 test had severely inflated size, as we expect the empirical size of the tests to be mostly within $1.96\sqrt{0.05 \cdot 0.95/1000} = .0135$ of the nominal level $\alpha = .05$. Additional simulation results not reported here show that the ratios of Kaplan-Meier cumulative distribution functions also resulted in substantial size inflation. For the ratios of Kaplan-Meier survivor functions, the size inflation was still a problem, though not as severe. All indications are that the tests based on ratios of single sample estimators need to be further refined before they can be recommended.

To compare the tests LRAD, LRAD2 targeting at improving the log rank test, we see that the LRAD test was more powerful than the logrank test across all configurations, as indicated by the power ratio staying greater than one. However, for small samples, the size of the LRAD test was inflated, although not as severely as for the CH and CH2 tests. The adjusted LRAD2 test brought the size under control. Most of the time it improved the power of the log rank test. For the few cases when it did not, the loss of power was virtually ignorable.

For comparing the MX test and the MXAD test, we see that the MXAD test improved the power over the MX test most of the time. Also, the magnitude of the relative power loss of the MXAD test for the few times where the MX test won out is mild. Thus the MXAD test can be recommended as an improvement of the MX test. Similar remarks also apply to the comparison between the MXB test and the MXBAD test, where the advantage of the MXBAD test can be seen for an even wider range of alternative hypotheses. Also, we note that when the treatment effect is initially adverse but becomes beneficial in the long run (case (S−L+)), most tests had much better performance than the logrank test, especially under heavy censoring, where the logrank test had substantial power loss. This is in agreement with Breslow et al. (1984) and with Yang and Prentice (2005), where the crossing hazards situations motivate the test and modeling respectively. To compare the MXAD, MXBAD, and the χ^2 tests, we note that the MXAD test often performed the best. The χ^2 test was often the least performing among the three robust tests. Under substantial censoring, it even had lower power than the logrank test most of the time. Also, the loss of power at the proportional alternatives can be quite sizable sometimes. This behavior indicates that the χ^2 test is in general not a good omnibus test when a range of alternatives is likely.

For comparing LRAD2 with the other tests, we see that, more often than not, the LAD2 test has higher power. It is the clear winner for the proportional hazards alternatives. For other scenarios, in situations where the LRAD2 test has less power compared with one of the

other robust tests, the LRAD2 test already has a sizable improvement over the logrank test. Due to its optimality at the proportional hazards alternatives, the improvement over the logrank test and consistent behavior across a range of nonproportional alternatives, we recommend the LRAD2 test as a good omnibus test. On the other hand, the MXBAD test would be a good choice for crossing hazards situations, and the MXAD test would generally be a good choice for other nonproportional hazards situations.

To study test robustness when the model of Yang and Prentice (2005) does not hold, we have conducted additional numerical simulations. Note that if (λ_C, λ_T) follow the model of Yang and Prentice (2005), then (λ_T, λ_C) does not. However, if the model of Yang and Prentice (2005) is fitted to (λ_T, λ_C), in general the resulting parameter estimate $\hat{\beta}$ is close to the negative of the true β for the (λ_C, λ_T) model. A few additional numerical simulations show that the tests behaved similarly to the results in Table 2, when the two groups were switched, and when data were generated for scenarios considered here but not under the model of Yang and Prentice (2005). The results are omitted here.

In the results reported here, we have focused on the robust combination tests. Single weighted logrank tests have also been examined. In general they behave as expected. Similar to the logrank test, they do well in a specific case (the weight S_C being optimal for (S+L0) and (S−L0), the proportional odds model), but are not as robust across several scenarios as those combination tests considered here.

5. Examples

We now illustrate these tests in several real data examples, ranging from moderate to severe deviations from the proportional hazards.

Example 1: Diverging survival functions

Nahman et al. (1992) studied the time to first exit-site infection (in months) in patients with renal insufficiency. In that study, 43 patients utilized a surgically placed catheter (Group 1) and 76 patients utilized a percutaneous placement of their catheter (Group 2). For testing H_0 with this data set, the logrank statistic is 1.599, resulting in the p-value of 0.110. Klein and Moeschberger (1997, Ch.7, Table 7.3) showed that a few other commonly used weighted logrank tests, with Gehan, Tarone-Ware, and Peto-Prentice weights, resulted in even larger p-values, while some weights, from the Fleming-Harrington $G^{\rho,\gamma}$ family that put more weight on the late comparisons, resulted in p-values less than .01. The plot of the estimated survival functions (Figure 7.1, Klein and Moeschberger, 1997) indicates that the two survival functions diverge at the right tail. This explains that weights that emphasize on late comparisons would likely yield small p-values. Table 3 gives the p-values of various tests. The LRAD test and the LRAD2 test use adaptive weights, so we do not need to decide which weight to use. The weight used in the MX, MXAD, MXB, and MXBAD tests is the Nelson-Aalen cumulative hazard estimator under H_0. Thus it is not surprising to see that they have p-values similar to those of weighted logrank tests with weights that emphasize on late comparisons. In this example, all six tests considered have more extreme p-values than the logrank test. The nonproportionality of hazard functions in this example is severe, and the MX and MXAD tests are better able to detect this nonproportionality, partly because of the particular weight used.

Example 2: Crossing survival functions

The Gastrointestinal Tumor Study Group (1982) compared chemotherapy with combined chemotherapy and radiation therapy, in the treatment of locally unresectable gastric cancer. Each treatment arm had 45 patients, with two observations of the chemotherapy group and six of the combination group censored. Kaplan-Meier plots of the two estimated survival

curves cross at around 1000 days. The logrank statistic for H_0 has the value 0.47, giving the p-value of 0.64. From Table 3, all tests except the MX test have p-values less that .05. This example indicates that the logrank test and even the MX test may not be sensitive enough to the crossing survival curves situation. With the use of adaptive weights, the MXAD test is able to correct the deficiency, resulting in a significant p-value. The MXB and the MXBAD tests are designed for this kind of situations and perform well here. The adaptive weights also lead to LRAD and LRAD2 tests with significant p-values.

In both of the examples above, there seems to be an apparent violation of proportional hazards. Not surprisingly, the logrank test behaved poorly and most other tests had much smaller p-values. For large clinical trials, often the deviation from the proportional hazards is less extreme. Next we look at such an example.

Example 3: Moderately time dependent hazard ratio

The Digoxin Intervention trial (The Digitalis investigation group, 1997) is a randomized, double-blind clinical trial on the effect of digoxin on mortality and hospitalization. In the main trial, patients with left ventricular ejection fraction of 0.45 or less were randomized to digoxin (3397 patients) or placebo (3403 patients) in addition to diuretics and angiotensin-converting-enzyme inhibitors. For testing the validity of the proportional hazards model, the acceleration test statistic of Breslow et al. (1984) has the value 1.6540, giving the p-value 0.098. Thus there is some indication of proportionality violation, but the evidence is not strong. The hazard ratio, obtained from fitting the model of Yang and Prentice (2005), varies mildly from .85 to .95. For the risk of mortality due to worsening heart failure, the logrank statistic has the value 1.88, resulting in the p-value of 0.061. From Table 3, none of the p-values are less than 0.05, but they are mostly less than 0.10. Thus there is some indication of lower risk of death attributed to worsening heart failure in the digoxin group, consistent with the trial's finding. This example shows that, for the case of moderately time-dependent hazard ratio, most of the robust modifications of the logrank test may reduce rather than boost the power. By contrast, the adaptively weighted logrank tests still provide more extreme p-values than the logrank test, although not as different as in the previous examples.

The above examples and simulation results, as well as additional simulation studies not reported here, show that typically the newly proposed tests improve respectively the tests that they are designed to modify. Overall, the LRAD2 test, the adjusted adaptively weighted logrank test, has the best performance. It improves the logrank test for almost all of the cases we have studied, and when it does not, has virtually ignorable power loss compared with the logrank test. On the other hand, if more information is available on the behavior of the two comparison groups, such as crossing survival curves, divergence or convergence survival curves, then weighted logrank test or appropriate robust tests such as the MX, MXAD, MXB, MXBAD tests, can be used to exploit the particular type of nonproportionality and improve the power.

6. Discussion

The statistics U_{Φ_1}, U_{Φ_2} are asymptotically equivalent to the logrank test statistic under H_0 and under the proportional hazards alternatives. This property gives the optimality for the adaptively weighted logrank test under proportional hazards, as well as the asymptotically correct size. Such property is likely to be useful in more situations than illustrated here. For example, they can also be used in the Renyi type test (Gill, 1980) where the maximum is taken over time. All the newly proposed tests can be modified appropriately to obtain the one-sided versions.

We have focused on the case of two homogeneous groups in this work. If this is not the case and covariates are available, then the covariates can be incorporated, just as for the weighted logrank statistics. The model of Yang and Prentice (2005) can be extended to accommodate covariates as indicated in their Discussion Section, and such an extension can be used for obtaining the adaptive weights.

Without proper adjustment, use of asymptotic approximations with adaptive weights often results in inflated size. We have studied a few cases of adjustment here that have worked well. Other options can be developed. Also, Monte Carlo methods, such as the Bootstrap or other resampling methods like that of Lin, Wei, and Ying (1993), can be used to obtain the *p*-values. The disadvantage is the increasing computing cost.

It is possible to consider other adaptive weights, and other models such as (20), particularly if additional hazard ratio patterns warrant them. For example, richer models that allow quadratic or *U*- shaped hazard ratios, or even more complicated patterns, can be considered. The drawback is the increasing complexity with more parameters to be estimated.

For design and analysis of clinical trials, sample size calculations and obtaining stopping boundaries for a sequential design are two important issues that need to be addressed. Incorporating the adaptive weights in extending the current clinical trial literature presents a very challenging but promising topic. All the issues arising from these various considerations remain to be explored in the future.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors are grateful to Editor Dr. David M. Zucker, the associate editor and the referee for their helpful comments that led to improvement of the manuscript. The research of Ross Prentice is supported by NIH grant CA 53996.

REFERENCES

Bennett S. Analysis of survival data by the proportional odds model. Statistics in Medicine. 1983; 2:273–277. [PubMed: 6648142]

Bickel, PJ.; Klaassen, CA.; Ritov, Y.; Wellner, JA. Efficient and Adaptive Estimation for Semiparametric Models. Baltimore, MD: Johns Hopkins Univ. Press; 1993.

Breslow N, Elder L, Berger L. A two sample censored-data rank test for acceleration. Biometrics. 1984; 40:1042–1069.

Cox DR. Regression models and life-tables (with Discussion). J. R. Statist. Soc. B. 1972; 34:187–220.

Fleming, TR.; Harrington, DP. Topics in Applied Statistics. New York: Marcel Dekker; 1984. Evaluation of censored survival data test procedures based on single and multiple statistics; p. 97-123.

Fleming, TR.; Harrington, DP. Counting Processes and Survival Analysis. New York: Wiley; 1991.

Fleming TR, Harrington DP, O'Sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. Journal of the American Statistical Association. 1987; 82:312–320.

Schein PD, Bruckner HW, Douglass HO, Mayer R, et al. GASTROINTESTINAL TUMOR STUDY GROUP. A comparison of combination chemotherapy and combined modality therapy for locally advanced gastric carcinoma. Cancer. 1982; 49:1771–1777. [PubMed: 6176313]

Gastwirth JL. The use of maximum efficiency robust tests in combining contingency tables and survival analysis. Journal of the American Statistical Association. 1985; 80:380–384.

Gill, R. Censoring and Stochastic Integrals. Math. Centre tract. Vol. 124. Amsterdam: Math. Centrum; 1980.

Kaplan E, Meier P. Nonparametric estimation from incomplete observations. Journal of the American Statistical Association. 1958; 53:457–481.

Klein, JP.; Moeschberger, ML. Survival Analysis: Techniques for Censored and Truncated Data. New York: Springer-Verlag; 1997.

Kosorok MR, Lin CY. The versatility of function-indexed weighted log-rank statistics J. Journal of the American Statistical Association. 1999; 94:320–332.

Lagakos SW, Lim LL-Y, Robins JM. Adjusting for early treatment termination in comparative clinical trials. Statistics in Medicine. 1990; 9:1417–1424. [PubMed: 2281229]

Lai TZ, Ying Z. Rank regression methods for left-truncated and right-censored data. Annals of Statistics. 1991; 19:531–556.

Lee JW. Some versatile tests based on the simultaneous use of weighted log-rank statistics. Biometrics. 1996; 52:721–725.

Lin DY, Wei LJ, Ying Z. Checking the Cox model with cumulative sums of martingale-based residuals. Biometrika. 1993; 80:557–572.

Mantel N. Evaluations of survival data and two new rank order statistics arising in its consideration. Caner Chemother. Rep. 1966; 50:163–170.

Nahman NS, Middendorf DF, Bay WH, Mcelligott R, Powell S, Anderson J. Modification of the percutaneous approach to peritoneal dialysis catheter placement under peritoneoscopic visualization: Clinical results in 78 patients. Journal of The American Society of Nephrology. 1992; 3:103–107. [PubMed: 1391701]

Peckova M, Fleming TR. Adaptive test for testing the difference in survival distributions. Lifetime Data Analysis. 2003; 9:223–238. [PubMed: 14649843]

Peto R, Peto J. Asymptotically efficient rank invariant test procedures (with Discussion). Journal of the Royal Statistical Society, Series A. 1972; 135:185–206.

Pollard, D. Empirical Processes: Theory and Applications. Hayward, CA: Institute of Mathematical Statistics; 1990.

Prentice RL. Linear rank tests with right censored data. Biometrika. 1978; 65:167–179.

Prentice RL, Pettinger M, Anderson GL. Statistical issues arising in the women's health initiative. Biometrics. 2005; 61:899–941. [PubMed: 16401257]

Self SG. An adaptive weighted log-rank test with application to cancer prevention and screening trials. Biometrics. 1991; 47:975–986. [PubMed: 1742450]

Shorack, GR.; Wellner, JA. Empirical Processes with Applications to Statistics. New York: Wiley; 1986.

Tarone RE. On the distribution of the maximum of the logrank statistic and the modified Wilcoxon statistic. Biometrics. 1981; 37:79–85.

Yang S, Hsu L, Zhao L. Combining asymptotically normal tests: cases studies in comparison of two groups. Journal of Statistical Planning and Inference. 2005; 133:139–158.

Yang S, Prentice RL. Semiparametric analysis of short term and long term relative risks with two sample survival data. Biometrika. 2005; 92:1–17.

Zucker DM. The efficiency of a weighted log-rank test under a percent error misspecification model for the log hazard ratio. Biometrics. 1992; 48:893–899. [PubMed: 1420846]

Zucker DM, Lakatos E. Weighted log rank type statistics for comparing survival curves when there is a time lag in the effectiveness of treatment. Biometrika. 1990; 77:853–864.

Table 1

Critical values for the modified Breslow et al. test with α = .05

$\tilde{\rho}$	1	.99	.95	.9	.8	.7	.5	.3	.1
$c_{\tilde{\rho}}(1-\sqrt{1-\alpha})$	2.235	2.29	2.346	2.381	2.422	2.446	2.473	2.485	2.49

Table 2

Size of the tests, and power ratio over the logrank test of these tests, with various sample sizes and censoring proportions, based on 1000 simulations.

(a) $n_1 = n_2 = 40$, 10% censoring

	CH	CH2	LRAD	LRAD2	MX	MXAD	MXB	MXBAD	χ^2
N	0.2550	0.1780	0.0910	0.0550	0.0580	0.0620	0.0580	0.0720	0.0550
PR+	1.6192	1.4136	1.1495	1.0421	0.9346	1.0140	0.7874	0.8224	0.8248
PR−	1.6514	1.4663	1.1514	1.0240	0.9183	0.9591	0.7668	0.8558	0.7740
S0L+	1.7531	1.5586	1.2943	1.1347	1.1721	1.1771	0.9401	1.0224	1.0299
S0L−	1.9278	1.7139	1.4588	1.2887	1.3995	1.3634	1.0876	1.2062	1.2088
S+L0	1.5938	1.4040	1.1722	1.0508	0.8190	1.0000	0.8499	0.9051	0.8852
S−L0	1.6731	1.5206	1.2397	1.1525	0.8692	1.0823	0.8886	1.0097	0.9443
S−L+	2.2868	2.1845	2.2768	2.1471	2.1197	2.2170	2.1820	2.3566	2.2668
S+L−	1.7722	1.6291	1.4881	1.3015	0.8807	1.3102	1.2863	1.4078	1.4100
U	1.9237	1.7099	1.4122	1.2468	0.8372	1.2010	0.8244	1.1170	0.8473

(b) $n_1 = n_2 = 40$, 50% censoring

	CH	CH2	LRAD	LRAD2	MX	MXAD	MXB	MXBAD	χ^2
N	0.2260	0.1480	0.0810	0.0580	0.0540	0.0540	0.0570	0.0530	0.0550
PR+	2.0352	1.6652	1.1674	0.9956	0.9648	1.0132	0.8722	0.8546	0.8590
PR−	1.8764	1.5491	1.1091	0.9927	0.8945	0.9418	0.7527	0.7891	0.7236
S0L+	2.5349	1.8760	1.3023	1.0465	1.1085	1.1008	0.9147	0.9225	0.8682
S0L−	3.2059	2.4020	1.3235	1.0882	1.0490	1.1373	0.9706	1.1078	0.9706
S+L0	1.6518	1.4005	1.1466	0.9921	0.8953	0.9634	0.7984	0.8115	0.8377
S−L0	1.6053	1.4189	1.2010	1.0678	0.8789	0.9903	0.8644	0.8838	0.8208
S−L+	7.7500	5.3036	8.4821	6.1250	2.4821	6.4821	7.9643	10.3929	7.8929
S+L−	1.2613	1.1446	1.0886	0.9969	0.9176	0.9767	0.8538	0.8507	0.8554
U	1.6224	1.4709	1.2890	1.1585	0.8625	1.1072	1.1352	1.1538	1.2541

(c) $n_1 = n_2 = 80$, 10% censoring

	CH	CH2	LRAD	LRAD2	MX	MXAD	MXB	MXBAD	χ^2
N	0.1820	0.1290	0.0580	0.0410	0.0380	0.0410	0.0530	0.0530	0.0400
PR+	1.2878	1.2267	1.0858	1.0334	0.9753	0.9971	0.8779	0.8997	0.8779
PR−	1.2518	1.1892	1.0612	1.0256	0.9616	0.9915	0.8634	0.8947	0.8734
S0L+	1.3261	1.2606	1.1601	1.0932	1.1325	1.1179	0.9520	1.0087	1.0175
S0L−	1.3271	1.2808	1.2214	1.1621	1.2171	1.2084	1.0434	1.1259	1.1505
S+L0	1.2718	1.2366	1.1338	1.0789	0.9141	1.0254	0.9225	0.9648	0.9563
S−L0	1.3023	1.2493	1.1318	1.0917	0.9155	1.0444	0.9427	1.0014	0.9771
S−L+	1.4807	1.4777	1.4792	1.4763	1.4763	1.4792	1.4733	1.4807	1.4822
S+L−	1.3469	1.3357	1.2993	1.2448	0.9245	1.2392	1.2196	1.2713	1.3035
U	1.3506	1.3147	1.1911	1.1365	0.9037	1.0991	0.8750	1.0302	0.8635

(d) $n_1 = n_2 = 80$, 50% censoring

	CH	CH2	LRAD	LRAD2	MX	MXAD	MXB	MXBAD	χ^2
N	0.2040	0.1460	0.0550	0.0470	0.0380	0.0390	0.0440	0.0470	0.0400
PR+	1.6726	1.4772	1.1193	1.0228	0.9365	0.9772	0.7868	0.7868	0.7868
PR−	1.5886	1.4464	1.1160	1.0438	0.9606	0.9847	0.8359	0.8643	0.8074
S0L+	2.4121	1.9286	1.3242	1.1099	1.2527	1.2253	0.8901	1.0330	0.9725
S0L−	2.4260	1.9704	1.2071	1.0355	1.1479	1.1538	0.9408	0.9941	0.9112
S+L0	1.3245	1.2429	1.0690	0.9984	0.9044	0.9592	0.8636	0.8652	0.8401
S−L0	1.2403	1.1861	1.0833	1.0403	0.8986	0.9986	0.8806	0.9222	0.8819
S−L+	11.4500	9.2333	11.7167	10.0000	4.2833	10.4333	13.3000	14.3333	13.2500
S+L−	1.0714	1.0593	1.0241	1.0011	0.9627	0.9824	0.9396	0.9418	0.9440
U	1.2921	1.2486	1.2008	1.1334	0.9031	1.1053	1.0913	1.1278	1.1882

Table 3

P- values of the tests for Examples 1 through 3

	LRAD	LRAD2	MX	MXAD	MXB	MXBAD
Example 1	0.034	0.047	0.006	0.007	0.016	0.013
Example 2	0.015	0.030	0.113	0.030	< 0.01	< 0.01
Example 3	0.051	0.055	0.091	0.082	0.120	0.107