Harvard University
Harvard University Biostatistics Working Paper Series

Year 2003 *Paper* 1

Nonparametric Comparison of Two Survival-Time Distributions in the Presence of Dependent Censoring

Greg DiRienzo*

*Harvard University, greg.dirienzo@gmail.com

Biometrics 59, 497–504
September 2003

Nonparametric Comparison of Two Survival-Time Distributions in the Presence of Dependent Censoring

A. G. DiRienzo

Department of Biostatistics, Harvard School of Public Health, Boston,
Massachusetts 02115, U.S.A.
email: dirienzo@biostat.harvard.edu

Summary. When testing the null hypothesis that treatment arm-specific survival-time distributions are equal, the log-rank test is asymptotically valid when the distribution of time to censoring is conditionally independent of randomized treatment group given survival time. We introduce a test of the null hypothesis for use when the distribution of time to censoring depends on treatment group and survival time. This test does not make any assumptions regarding independence of censoring time and survival time. Asymptotic validity of this test only requires a consistent estimate of the conditional probability that the survival event is observed given both treatment group and that the survival event occurred before the time of analysis. However, by not making unverifiable assumptions about the data-generating mechanism, there exists a set of possible values of corresponding sample-mean estimates of these probabilities that are consistent with the observed data. Over this subset of the unit square, the proposed test can be calculated and a rejection region identified. A decision on the null that considers uncertainty because of censoring that may depend on treatment group and survival time can then be directly made. We also present a generalized log-rank test that enables us to provide conditions under which the ordinary log-rank test is asymptotically valid. This generalized test can also be used for testing the null hypothesis when the distribution of censoring depends on treatment group and survival time. However, use of this test requires semiparametric modeling assumptions. A simulation study and an example using a recent AIDS clinical trial are provided.

Key words: Log-rank test; Randomized clinical trial; Sensitivity analysis; Survival analysis.

1. Introduction

1.1 *Background*

Consider a randomized clinical trial that enrolls patients through time and follows them for the occurrence of a primary event. At the time of analysis, the time from randomization to the primary event, or survival time, is potentially censored by the administrative censoring time, the calendar date of analysis minus the calendar date of randomization. At the time of analysis, this potential administrative censoring time is observed in full for each subject enrolled in the study and, because of randomization, its distribution is conditionally independent of treatment group given survival time. However, the time to primary event can be censored before the administrative censoring time, for example, when a patient is lost to follow-up before the time of analysis. The time from randomization to an event, other than the time of analysis, that can censor the survival time is referred to as a "nonadministrative censoring time." The actual censoring time for each subject is then the minimum of the administrative censoring time and the nonadministrative censoring time; this actual censoring time will henceforth be referred to simply as the censoring time. Unlike the administrative censoring time, the censoring time is not necessarily observed in full for each subject at the time of the analysis. For example, when the primary event is observed and this event is death, the censoring event is only known to lie in the interval between death and analysis time.

Study 320 of the AIDS Clinical Trials Group, ACTG 320, enrolled patients from January 1996 to January 1997 (Hammer et al., 1997). The patients were randomized to receive either the drug combination ZDV+3TC+placebo (581 patients) or ZDV+3TC+indinavir (575 patients). The primary endpoint of the study was the time from randomization to either death or AIDS, whichever occurred first. On February 18, 1997, an interim analysis was conducted that assumed noninformative censoring, and the decision was made to stop accrual and close the study, because of a significant beneficial effect of indinavir. The assumption of noninformative censoring here states that the distribution of censoring time is conditionally independent of the time to either death or AIDS, whichever comes first, given treatment group and covariates. There were 66 primary events observed in the placebo group and 38 primary events observed in the indinavir group. For each patient, the administrative censoring time was the calendar date February 18, 1997, minus the calendar date of enrollment; the median administrative censoring time was 293 days in each treatment group. The number of patients who were lost to follow-up before their respective administrative censoring time was 55 in the placebo group and 39 in the indinavir group.

1.2 *Preview of Results*

In this article, we show that under the null hypothesis that the survival-time distribution is independent of randomized treatment group, the log-rank test is asymptotically valid when the distribution of the time to censoring is conditionally independent of treatment group given survival time. This is true regardless of whether the censoring time distribution depends on survival time. An example of when the censoring time distribution is conditionally independent of randomized treatment group given survival time, but not independent of survival time, is when the only form of censoring is administrative and a trend in patient accrual exists. An example of this trend is when sicker patients (with shorter survival times on average) tend to enroll later in the study and as a result have shorter censoring times on average than those healthier patients (with longer survival times on average) who tended to enroll prior to them and thus have on average longer censoring times. On the other hand, when the distribution of censoring time depends on treatment group given survival time, the log-rank test is asymptotically valid when the distribution of censoring time is conditionally independent of survival time given treatment group. An example of such a setting occurs when no trend in patient accrual exists and the only competing cause of censoring is from study dropout because of a toxicity that is more likely to occur in one treatment group than in another, but the distribution of dropout time associated with this toxicity is conditionally independent of survival time given treatment group.

For the case when the distribution of censoring depends on treatment group and survival time, we derive a two-sample test of the null hypothesis. This test requires a consistent estimate of the conditional probability that, under the null hypothesis, the survival event is observed given treatment group and that the survival event occurred before the time of analysis; for ease of exposition, denote these two probabilities here by p_0 and p_1 and their respective sample-mean estimates by $\hat{p}_0$ and $\hat{p}_1$. Unfortunately, without making unverifiable assumptions about the data-generating mechanism, there exists a set of possible values of $\hat{p}_0$ and $\hat{p}_1$ that are consistent with the observed data. Thus, in order to properly execute such testing methodology, the proposed test needs to be calculated over this corresponding subset of the unit square and the region where the test rejects must be identified. Subject-matter experts may then be elicited for judgments about plausible ranges for these probability estimates. A decision on the null hypothesis can then be made after a pure quantification of uncertainty about this decision because of possible dependence between censoring time, survival time, and treatment group. That is to say, such a quantification of uncertainty does not rely on unverifiable assumptions about the data-generating mechanism. Although the ultimate decision on the null hypothesis is based on a subjective decision on the plausible range for $\hat{p}_0$ and $\hat{p}_1$, one can directly assess the effect of this subjectivity on their decision.

The probability $1 - p_j$ may be easier to interpret than p_j, $j = 0, 1$. The quantity $1 - p_j$ denotes the conditional probability under the null hypothesis of, for a subject in treatment group j, not observing the survival event given that it would have been observed had the only form of censoring been administrative. Eliciting information about plausible ranges for values of $1 - \hat{p}_j$ should thus proceed by first thinking about the subgroup of patients in the study for which the survival event occurs before the analysis time under the null hypothesis. Then, plausible ranges for the likelihood of not observing the survival event in this subgroup because of competing causes of censoring need to be decided on for each treatment group. Referring back to ACTG 320, an interesting question to ask is, "For what ranges of values for $1 - \hat{p}_0$ and $1 - \hat{p}_1$ would the decision to declare indinavir superior be overturned, and, would such ranges be considered plausible?" The reader should not be discouraged by the fact that $\hat{p}_0$ and $\hat{p}_1$ are not uniquely identified from the observed data. Often, precise ranges for values of these estimates are not necessary to obtain a decision on the null hypothesis. For instance, in ACTG 320, the range of values for $1 - \hat{p}_0$ and $1 - \hat{p}_1$ that are consistent with the decision to declare indinavir superior is given by $1 - \hat{p}_0 \geq 1 - \hat{p}_1$.

An attractive feature of this newly introduced test is that only two scalar sensitivity parameters, representing the unknown values $\hat{p}_0$ and $\hat{p}_1$, are required, each with range within the unit interval. The testing methodology we propose exploits this feature by recognizing that a complete sensitivity analysis of the test can directly proceed without having to make unverifiable modeling assumptions to make sensitivity analyses feasible by reducing the dimensionality of the required sensitivity parameters. For example, dimension reduction of required sensitivity parameters was necessary for the sensitivity methodology proposed by Scharfstein, Rotnitzky, and Robins (1999); this was because their methods concerned estimation of the mean of a continuous variable in the presence of dependent censoring.

We also generalize the log-rank test for use in the setting where the distribution of time to censoring depends on survival time and treatment group; in turn, this enables us to provide conditions under which the ordinary log-rank test is asymptotically valid, which were stated above. The asymptotic validity of this generalized test requires a consistent estimate of an infinite-dimensional parameter within each treatment group. Unfortunately, the corresponding sample-mean estimate of this probability is not uniquely identified from the observed data. In order for inference to proceed with this test, unverifiable semiparametric modeling assumptions are required whose relevant parameters are not able to be estimated from the observed data; a sensitivity analysis involving these parameters is thus required. Although it is certainly possible for a data analyst to perform such semiparametric modeling and conduct associated sensitivity analyses, apart from identifying some possible semiparametric models, we do not expand this area, as it is not an aim of this article.

DiRienzo and Lagakos (2001a) propose a class of two-sample tests that can be used when the distribution of censoring depends on treatment group and survival time. However, for these tests to be asymptotically valid, it is required that (i) the unverifiable assumption that the times to censoring and survival are conditionally independent given treatment group and covariates holds, and (ii) either the conditional distribution of time to censoring given treatment group and covariates or the conditional distribution of survival time given treatment group and covariates is correctly modeled. The test proposed in this article essentially replaces unverifiable assumptions with a sensitivity analysis, where one can directly examine the pure impact of dependent censoring on a decision on the null hypothesis.

Some relevant literature concerns estimation of the survival function of a continuous failure-time variable in the presence of dependent censoring. This literature can be separated into methods that do incorporate information from variables related to survival time and censoring, so-called auxiliary variables, and methods that do not. Estimation techniques that make the nonidentifiable assumption that information from all such auxiliary variables is available are given by Robins and Rotnitzky (1992), Robins (1993), Robins and Finkelstein (2000), and Satten, Datta, and Robins (2001). When information from auxiliary variables is not available, nonidentifiable assumptions need to be made about the dependence structure between survival time and censoring time. Methods that vary such assumptions in a sensitivity analysis are provided by Fisher and Kanarek (1974), Slud and Rubinstein (1983), Klein and Moeschberger (1988), Klein et al. (1992), Moeschberger and Klein (1995), and Zheng and Klein (1995). Recently, Scharfstein and Robins (2002) have presented methods for estimating the survival function that assumes some but not all auxiliary variables are available, and proposed methods of analysis that investigate the sensitivity of inference to residual dependence between survival and censoring due to unmeasured auxiliary variables.

This article is organized as follows. In Section 2 we define test statistics and present methods for inference. Section 3 provides simulation results and Section 4 illustrates the methodology on a recent AIDS clinical trial.

2. Test Statistics and Inference

2.1 *Notation*

Let the binary random variable R denote treatment group and let W denote a vector of baseline covariates. Information from covariates can be used in the methods we propose, by defining strata within which the proposed testing procedures can be conducted. Let the continuous random variable X denote time from randomization to the primary event and let C be the potential administrative censoring time. Note that C is observed in full for each subject at the time of analysis. We work within the context of a randomized clinical trial with possibly staggered entry, and thus assume that the conditions $R \perp W$ (i.e., the distribution of covariates is independent of treatment group) and $C \perp R \mid X$ hold throughout this article. Note that the distribution of C is allowed to depend on X, which would be the case when a trend in patient accrual exists, for example, when sicker patients tend to enter the study later than more healthy ones. Let $X^* = \min(X, C)$ and $\delta^C = I(X \leq C)$, where $I(\cdot)$ is the indicator function. Let D denote the time from randomization to an event other than time of analysis that censors X; for example, D may be the time from randomization to loss to follow-up. The actual censoring time is thus $C^* = \min(C, D)$. Denote $\tilde{X} = \min(X, C, D)$ and $\delta = I(\tilde{X} = X)$. Note that δ is always observed, but when $\delta = 0$ and $\tilde{X} < C$, δ^C is missing. The data is assumed to consist of n independent and identically distributed realizations of $(R, W, C, \tilde{X}, \delta)$, denoted by $(R_i, W_i, C_i, \tilde{X}_i, \delta_i)$, $i = 1, \ldots, n$. No other assumptions about the data-generating mechanism are made in this article, including the frequently made assumption that censoring acts noninformatively, that is, $X \perp C^* \mid (R, W)$.

2.2 *Test Statistic* $L_n(\rho)$

We consider tests of the null hypothesis $H_0 : R \perp X$, that the survival-time distribution does not depend on treatment group. Define the statistic

$$n^{-\frac{1}{2}} U_n(\rho) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(R_i)\delta_i\{R_i - E_n(R)\},$$

where the notation $E_n(Z)$ denotes the sample mean of the random variables $\{Z_1, \ldots, Z_n\}$ and $\rho(R_i)$ is defined below. It is straightforward to show that

$$n^{-\frac{1}{2}} U_n(\rho) = n^{-\frac{1}{2}} \sum_{i=1}^{n} A_i(\rho) + o_p(1),$$

where

$$A_i(\rho) = (R_i - \pi)[\rho(R_i)\delta_i - E\{\rho(R)\delta\}], \quad i = 1, \ldots, n,$$

are independent and identically distributed terms. To see this, note that

$$\begin{aligned} n^{-\frac{1}{2}} U_n(\rho) = n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(R_i)\delta_i\{R_i - E(R)\} \\ - n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(R_i)\delta_i\{E_n(R) - E(R)\} \end{aligned}$$

and the second term on the right-hand side in the line above can be written as

$$\begin{aligned} & n^{-\frac{1}{2}} \sum_{i=1}^{n} \{R_i - E(R)\} E\{\rho(R)\delta\} \\ & \quad + n^{-\frac{1}{2}} \sum_{i=1}^{n} \{R_i - E(R)\} \left[n^{-1} \sum_{i=1}^{n} \rho(R_i)\delta_i - E\{\rho(R)\delta\} \right]. \end{aligned}$$

By Slutzky's theorem, the second term in the line above is $o_p(1)$.

Using the fact $\mathrm{pr}(\delta = 1 \mid R) = \mathrm{pr}(\delta = 1, \delta^C = 1 \mid R)$, it follows that

$$\begin{aligned} E\{A(\rho)\} = E[\{R - E(R)\}\mathrm{pr}(\delta^C = 1 \mid R)\rho(R) \\ \times \mathrm{pr}(\delta = 1 \mid \delta^C = 1, R)]. \end{aligned}$$

Under H_0 and the condition $C \perp R \mid X$, which is satisfied in a randomized clinical trial and is assumed to hold throughout, $\mathrm{pr}(\delta^C = 1 \mid R) = \mathrm{pr}(\delta^C = 1)$ and, with the definition $\rho(R) = 1/\mathrm{pr}(\delta = 1 \mid \delta^C = 1, R)$, results in $E\{A(\rho)\} = \mathrm{pr}(\delta^C = 1)E\{R - E(R)\} = 0$. Here it is assumed that $\mathrm{pr}(\delta = 1 \mid \delta^C = 1, R)$ is bounded away from 0, i.e., $\mathrm{pr}(\delta = 1 \mid \delta^C = 1, R) \geq \epsilon > 0$, ϵ arbitrary. Therefore, under H_0, $n^{-\frac{1}{2}} U_n(\rho)$ is asymptotically normal with mean 0 and variance $\sigma^2(\rho) = E\{A^2(\rho)\}$. It can be shown that a consistent estimator of $\sigma^2(\rho)$ is $\sigma_n^2(\rho) = (1/n) \sum [A_i^{(n)}(\rho) - E_n\{A^{(n)}(\rho)\}]^2$, with

$$A_i^{(n)}(\rho) = \{R_i - E_n(R)\}[\rho(R_i)\delta_i - E_n\{\rho(R)\delta\}], \quad i = 1, \ldots, n.$$

Thus, an asymptotically valid test of H_0 is given by $L_n(\rho) = n^{-\frac{1}{2}}U_n(\rho)/\{\sigma_n^2(\rho)\}^{\frac{1}{2}}$.

Suppose that $n^{-\frac{1}{2}}U_n(\hat{\rho})$ was defined as $n^{-\frac{1}{2}}U_n(\rho)$ except that $\rho(R=0)$ and $\rho(R=1)$ are replaced by consistent point estimates. Then, by using a Taylor series expansion, it is easily shown that $n^{-\frac{1}{2}}U_n(\hat{\rho})$ and $n^{-\frac{1}{2}}U_n(\rho)$ have the same asymptotic distribution. Similarly, if $\sigma_n^2(\hat{\rho})$ is defined as $\sigma_n^2(\rho)$ except with $\rho(R)$ being replaced by consistent estimates, it can be shown that $\sigma_n^2(\hat{\rho})$ is also a consistent estimate of $\sigma^2(\rho)$. As a result, $L_n(\hat{\rho}) = n^{-\frac{1}{2}}U_n(\hat{\rho})/\{\sigma_n^2(\hat{\rho})\}^{\frac{1}{2}}$ is asymptotically standard normal under H_0.

The sample-mean estimate of $\text{pr}(\delta = 1 \,|\, \delta^C = 1,\, R)$, say, $\widehat{\text{pr}}(\delta = 1 \,|\, \delta^C = 1, R)$ is not uniquely identified from the observed data, since δ^C is missing when $\delta = 0$ and $\tilde{X} < C$. The following diagram of the observed data for a given treatment group illustrates this.

$$\begin{array}{r|c} \text{value of } \tilde{X} & \text{value of } \delta^C \\ \hline \delta = 1 \left\{ \begin{array}{c} X \\ \vdots \\ X \end{array} \right. & \begin{array}{c} 1 \\ \vdots \\ 1 \end{array} \\ \delta = 0 \left\{ C^* < C \left\{ \begin{array}{c} D \\ \vdots \\ D \end{array} \right. \right. & \begin{array}{c} ? \\ \vdots \\ ? \end{array} \\ \hline\hline \delta = 0 \left\{ C^* = C \left\{ \begin{array}{c} C \\ \vdots \\ C \end{array} \right. \right. & \begin{array}{c} 0 \\ \vdots \\ 0 \end{array} \end{array}$$

Note that the case when $\delta^C = 0$ does not play a role in estimation of $\text{pr}(\delta = 1 \,|\, \delta^C = 1,\, R)$.

Had δ^C been observed for all subjects, the set of possible values for $\widehat{\text{pr}}(\delta = 1 \mid \delta^C = 1, R)$ extends from the case corresponding to when all those with $\delta = 0$ and $\tilde{X} < C$ have $\delta^C = 0$ to the case corresponding to when all these subjects have $\delta^C = 1$, for which the sample-mean estimate equals 1. In notation, the set of possible values for $\hat{\text{pr}}(\delta = 1 \mid \delta^C = 1, R)$ is, for $R_i = 0$,

$$\left\{ \frac{\sum(1-R_i)\delta_i}{\sum(1-R_i)\{\delta_i + I(\tilde{X}_i < C_i, \delta_i = 0)\}}, \right. \\ \left. \frac{\sum(1-R_i)\delta_i}{\left[\sum(1-R_i)\{\delta_i + I(\tilde{X}_i < C_i, \delta_i = 0)\}\right] - 1}, \ldots, 1 \right\}$$

and for $R_i = 1$, is

$$\left\{ \frac{\sum R_i\delta_i}{\sum R_i\{\delta_i + I(\tilde{X}_i < C_i, \delta_i = 0)\}}, \right. \\ \left. \frac{\sum R_i\delta_i}{\left[\sum R_i\{\delta_i + I(\tilde{X}_i < C_i, \delta_i = 0)\}\right] - 1}, \ldots, 1 \right\}$$

$i = 1, \ldots, n$. The test $L_n(.)$ thus needs to be calculated over this grid of possible values for $\widehat{\text{pr}}(\delta = 1 \mid \delta^C = 1, R)$ and the rejection region identified. Plausible ranges for these probability estimates need to be obtained, if possible with aid from subject-matter experts, and a decision on H_0 made. Note that the value of $L_n(.)$ on the identity line within $[\epsilon, 1] \times [\epsilon, 1]$ makes no correction for dependent censoring.

The test $L_n(\hat{\rho})$ estimates the value of the test $L_n(\hat{\rho}), \hat{\rho} = 1$, that would have arisen had, contrary to fact, the only form of censoring been administrative censoring. This is because, when the only form of censoring is administrative, δ_i^C is observed for each subject $i = 1, \ldots, n$, and $\widehat{\text{pr}}(\delta = 1 \mid \delta^C = 1, R) = 1$. Given the observed data, the range of possible values for $L_n(\hat{\rho}), \hat{\rho} = 1$, that could have arisen had the only form of censoring been administrative can be obtained by imputing values of δ_i^C for those subjects for which δ_i^C is missing in such a way that results in the corresponding largest (and smallest) possible value that could have been observed for $L_n(\hat{\rho}), \hat{\rho} = 1$. This possible range is denoted by $[\ell_n^{\min},\, \ell_n^{\max}]$. Here, $\ell_n^{\min}$ is obtained by imputing the values $\delta_i^C = 1$ (and $\delta_i = 1$) for those cases with $R_i = 0, \delta_i = 0, \tilde{X}_i < C_i$ (the cases in group $R_i = 0$ for which δ_i^C is missing), and leaving the observed values of δ_i unchanged for cases in group $R_i = 1$ (imputing $\delta_i^C = 0$ for those missing). Similarly, $\ell_n^{\max}$ is calculated by imputing the value $\delta_i^C = 1$ (and $\delta_i = 1$) for those cases with $R_i = 1, \delta_i = 0, \tilde{X}_i < C_i$ (the cases in group $R_i = 1$ for which δ_i^C is missing), and leaving the observed value of δ_i unchanged for cases in group $R_i = 0$ (imputing $\delta_i^C = 0$ for those missing). Sensitivity analyses should thus only consider values of $L_n(.)$ in $[\ell_n^{\min},\, \ell_n^{\max}]$.

2.3 *A Generalized Log-Rank Test*

A test that may be more efficient than $L_n(\rho)$ can be constructed from the rank statistic

$$n^{-\frac{1}{2}}\tilde{U}_n(\rho, \phi) \\ = n^{-\frac{1}{2}} \sum_{i=1}^{n} \rho(R_i) \int \left[R_i - \frac{E_n\{Y(x)\phi(x;R)R\}}{E_n\{Y(x)\phi(x;R)\}} \right] dN_i(x),$$

where $Y_i(x) = I(x \le \tilde{X}_i)$, $N_i(x) = I(x \ge \tilde{X}_i, \delta_i = 1)$, $i = 1, \ldots, n$ and $\phi(x;R_j)$ is defined below. Note that this statistic uses information about the times of failure, which may lead to a test that is more efficient than $L_n(\rho)$, which makes no use of the failure times. The probability limit of $E_n\{Y(x)\phi(x;R)R\}/E_n\{Y(x)\phi(x;R)\}$ is

$$\mu(x, \phi) = \frac{E\{R\text{pr}(X^* \ge x \mid R)\phi(x;R)\text{pr}(D \ge x \mid X^* \ge x, R)\}}{E\{\text{pr}(X^* \ge x \mid R)\phi(x;R)\text{pr}(D \ge x \mid X^* \ge x, R)\}}.$$

Under H_0 and the condition $C \perp R \,|\, X$, $\text{pr}(X^* \ge x \,|\, R) = \text{pr}(X^* \ge x)$ and, for example, with the definition $\phi(x;R) = 1/\text{pr}(D \ge x \,|\, X^* \ge x,\, R)$, then $\mu(x,\, \phi) = E(R)$; here, $\text{pr}(D \ge x \,|\, X^* \ge x,\, R)$ is assumed to be uniformly bounded away from 0, i.e., $\text{pr}(D \ge x \,|\, X^* \ge x,\, R) \ge \eta > 0$, $x > 0$, η arbitrary. Note that the choice $\phi(x; R = 1) = 1$ and $\phi(x; R = 0) = \alpha(x)$, with $\alpha(x) = \text{pr}(D \ge x \,|\, X^* \ge x, R = 1)/\text{pr}(D \ge x \,|\, X^* \ge x, R = 0)$ also results in $\mu(x,\, \phi) = E(R)$ here. This result implies that only the ratio $\alpha(x)$ needs to be estimated. To see this, note that under H_0 and $C \perp R \,|\, X$, $\mu(x,\, \phi)$ may in general be written as $\mu(x,\, \phi) = E(R)/[E(R) + \{1 - E(R)\}\psi(x)]$, where

$$\psi(x) = \{\phi(x; R = 0)\text{pr}(D \ge x \mid X^* \ge x, R = 0)\}/ \\ \{\phi(x; R = 1)\text{pr}(D \ge x \mid X^* \ge x, R = 1)\}.$$

It can be shown, using arguments similar to those in the appendix of DiRienzo and Lagakos (2001a), that

$$n^{-\frac{1}{2}}\tilde{U}_n(\rho,\phi) = n^{-\frac{1}{2}}\sum_{i=1}^{n} B_i(\rho,\phi) + o_p(1),$$

where

$$\begin{aligned} B_i(\rho,\phi) = \{R_i - E(R)\} \\ \times \int \Bigg[\rho(R_i)dN_i(x) - \frac{Y_i(x)\phi(x;R_i)}{E\{Y(x)\phi(x;R)\}} \\ \times E\{\rho(R)dN(x)\}\Bigg], \quad i = 1,\ldots,n, \end{aligned}$$

are independent and identically distributed terms that have mean 0 under H_0 for both choices of $\phi(.)$ given above. Therefore, under H_0, $n^{-\frac{1}{2}}\tilde{U}_n(\rho,\phi)$ is asymptotically normal with mean 0 and variance $\tilde{\sigma}^2(\rho,\phi) = E\{B^2(\rho,\phi)\}$. Again, using arguments similar to those in the appendix of DiRienzo and Lagakos (2001a), it can be shown that a consistent estimate of $\tilde{\sigma}^2(\rho,\phi)$ is $\tilde{\sigma}_n^2(\rho,\phi) = (1/n)\sum[B_i^{(n)}(\rho,\phi) - E_n\{B^{(n)}(\rho,\phi)\}]^2$, where $B_i^{(n)}(\rho,\ \phi)$ is defined as $B_i(\rho,\ \phi)$, except with E_n replacing E, $i = 1,\ldots,n$. An asymptotically valid test of H_0 is thus given by $\tilde{L}_n(\rho,\phi) = n^{-\frac{1}{2}}\tilde{U}_n(\rho,\phi)/\{\tilde{\sigma}_n^2(\rho,\phi)\}^{\frac{1}{2}}$.

When only administrative censoring is possible, $\rho(R) = 1$ and $\phi(x;R) = 1$, for $x > 0$, and $n^{-\frac{1}{2}}\tilde{U}_n(\rho,\phi)$ is the numerator of the ordinary log-rank test; this establishes the asymptotic validity of the log-rank test in this case. Note that the numerator of $\tilde{L}_n(\rho,\phi)$ equals that of the log-rank test whenever $C^* \perp R \,|\, X$. Also note that when the distribution of C^* depends on R given X, but the condition $C^* \perp X \,|\, R$ holds, it can be shown using techniques similar to those in DiRienzo and Lagakos (2001b) that the ordinary log-rank test is asymptotically valid.

When $\tilde{X} < x$, $\delta = 0$, and $C \geq x$, it is only known that $X^* \in (\tilde{X}, C]$ and thus the sample-mean estimate of $\text{pr}(D \geq x \,|\, X^* \geq x,\ R)$ is not uniquely identified from the observed data. However, given the observed data, the set of possible values for the sample-mean estimate of $\text{pr}(D \geq x \,|\, X^* \geq x,\ R = 1)$ had X^* and δ^C been observed for all subjects, begins with the point $\sum\{R_i I(\tilde{X}_i \geq x)\}/\sum R_i\{I(\tilde{X}_i \geq x) + I(\tilde{X}_i < x, \delta_i = 0, C_i \geq x)\}$ and continues by unit decrements in the denominator to 1, $i = 1,\ldots,n$; similarly for $R_i = 0$. Thus, there exists a set of possible values for the sample-mean estimate of $\alpha(x)$, say, $\hat{\alpha}(x)$, for $x > 0$. However, it is in general difficult to enumerate all possible values of $\hat{\alpha}(x)$, for $x > 0$; this is unlike the case for the parameter $\rho(R)$, for which all possible values for the sample-mean estimates are a subset of the unit square. One way around this problem is to specify a parametric model for $\hat{\alpha}(x)$, e.g., $\hat{\alpha}(x) = \hat{\alpha}$ or $\hat{\alpha}(x) = \exp(\hat{\alpha}x)$, where in both cases, $\hat{\alpha}$ is not able to be calculated from the observed data and needs to be treated as a sensitivity parameter. If the test $\tilde{L}_n(\hat{\rho},\hat{\phi})$ is defined as $\tilde{L}_n(\rho,\phi)$, except with $\rho(R)$ replaced by the corresponding sample-mean estimate and $\alpha(x)$ replaced by a correctly specified parametric model estimate, then under H_0, $\tilde{L}_n(\hat{\rho},\hat{\phi})$ is asymptotically standard normal.

One approach to proceed with sensitivity analyses is to condition on a choice for the pair $\widehat{\text{pr}}(\delta = 1 \mid \delta^C = 1, R)$, and vary $\hat{\alpha}(x)$ over a plausible range, repeating this for a range of choices for $\widehat{\text{pr}}(\delta = 1 \mid \delta^C = 1, R)$. Also, as with the test $L_n(\hat{\rho})$, it can be shown that there is a possible range for the test $\tilde{L}_n(\hat{\rho},\hat{\phi})$ given the observed data; only this range of values should be considered by sensitivity analyses. The details of this calculation are omitted here, but can be obtained on request from the author.

3. Simulation Study

To evaluate the small sample properties of the proposed test statistics, we conducted simulations of their behavior when the distribution of censoring was dependent on treatment group given survival time, as well as (i) dependent on survival time given treatment group and (ii) conditionally independent of survival time given treatment group. Case (ii) investigates whether there is a penalty for unnecessarily using the proposed tests, as, in this nonidentifiable setting, the ordinary log-rank test is asymptotically valid.

For each of 2000 independent simulation iterations: R was simulated via a random allocation design, with $E(R) = 1/2$ and W as a binary variable with $\text{pr}(W = 1) = 1/2$. The survival time was distributed as log-normal, with $\log X = -0.75W + \varepsilon$, where $\varepsilon \sim N(1,\ 0.5^2)$. The 25th, 50th, and 75th percentiles of this distribution of X are approximately 1, 2, and 3, respectively. The administrative censoring time C was simulated as $U(2, 4)$; this resulted in $\text{pr}(\delta^C = 1)$ being approximately equal to 3/4. For case (i), for those with $R = 0$, $D = C$, so that X could only be censored administratively; for those with $R = 1$, $D \sim U(0, 3)$ if $W = 0$ and $D \sim U(1, 4)$ if $W = 1$. This scheme resulted in $\text{pr}(\delta = 1)$ being approximately equal to 0.62 and $\text{pr}(\delta = 1 \,|\, \delta^C = 1,\ R = 1)$ being approximately equal to 2/3. Note that $\text{pr}(\delta = 1 \,|\, \delta^C = 1,\ R = 0) = 1$. The simulation results corresponding to this setting are shown in part (a) of Table 1. For case (ii), all variables were generated as above except D, which was taken as $D = C$ for those with $R = 0$ and, for those with $R = 1$, $D \sim U(1, 4)$; this resulted in $\text{pr}(\delta = 1)$ being approximately equal to 2/3 and $\text{pr}(\delta = 1 \,|\, \delta^C = 1, R = 1)$ being approximately equal to 0.78. The corresponding simulation results are presented in part (b) of Table 1. The average lower bound for the sample-mean estimate of $\text{pr}(\delta = 1 \,|\, \delta^C = 1,\ R = 1)$ given the observed data was approximately 0.5 for setting (i) and 0.65 for setting (ii). Finally, $\phi(x; R = 1)$ was equal to 1 and $\phi(x; R = 0) = \alpha(x)$ was estimated from an independent sample of n = 200,000 and then held fixed throughout.

In setting (i), the log-rank test has an empirical size well above the nominal level and the magnitude of this bias increases with sample size. However, at the true values of $\rho(R)$ and $\phi(x;R)$, the proposed tests reject near the nominal rate. Note that the performance of the proposed tests becomes worse the further values of $\hat{\rho}(R)$ are from $\rho(R)$. In setting (ii), the log-rank test rejects near the nominal rate as expected; however, the corrected tests reject at the nominal rate only near the true value $\rho(R)$.

Simulations were also conducted under H_0 for the case when the distribution of censoring was conditionally independent of treatment group given survival time, as well as (i) conditionally independent of survival time given treatment group, and (ii), dependent on survival time given treatment group. For both cases, all variables were simulated as above except that for case (i), $D \sim U(1, 4)$, and

Table 1

Empirical mean, standard error, and size of ordinary log-rank test and proposed tests for treatment effect at the 0.05 *nominal level;* $\mathrm{pr}(\delta = 1 \mid \delta^C = 1,\, R = 0) = 1$

(a)		$C^*\leftrightarrow R \mid X,\ C^*\leftrightarrow X \mid R$					
		$n = 200$			$n = 500$		
Test	$1/\hat\rho(R=1)$	Size	Mean	SE	Size	Mean	SE
Log-rank	–	0.165	0.97	1.00	0.304	1.45	1.01
$L_n(\hat\rho)$	0.5	0.593	2.21	1.00	0.936	3.45	1.02
$\tilde L_n(\hat\rho,\phi)$		0.397	1.69	1.03	0.671	2.45	1.04
$L_n(\rho)$	0.66	0.056	−0.08	1.01	0.057	−0.01	1.01
$\tilde L_n(\rho,\phi)$		0.068	0.23	1.04	0.064	0.14	1.04
$L_n(\hat\rho)$	0.83	0.569	−2.11	1.04	0.906	−3.37	1.05
$\tilde L_n(\hat\rho,\phi)$		0.163	−0.93	1.03	0.398	−1.68	1.03

(b)		$C^*\leftrightarrow R \mid X,\ C^* \perp X \mid R$					
		$n = 200$			$n = 500$		
Test	$1/\hat\rho(R=1)$	Size	Mean	SE	Size	Mean	SE
Log-rank	–	0.048	0.06	1.00	0.057	−0.01	1.03
$L_n(\hat\rho)$	0.65	0.416	1.76	1.03	0.753	2.68	1.02
$\tilde L_n(\hat\rho,\phi)$		0.199	1.11	1.00	0.377	1.64	1.02
$L_n(\rho)$	0.78	0.055	0.04	1.02	0.059	−0.04	1.02
$\tilde L_n(\rho,\phi)$		0.053	0.09	1.01	0.052	0.03	1.03
$L_n(\hat\rho)$	0.9	0.327	−1.40	1.03	0.628	−2.31	1.03
$\tilde L_n(\hat\rho,\phi)$		0.108	−0.72	1.01	0.239	−1.25	1.03

for case (ii), $D \sim U(0,\, 3)$ if $W = 0$ and $D \sim U(1,\, 4)$ if $W = 1$. For $n = 200$, the empirical size, mean, and standard error of $L_n(1)$ and $\tilde L_n(1,1)$ were (0.059, −0.01, 1.04) and (0.058, −0.02, 1.05), respectively, for case (i), and (0.055, −0.01, 1.03) and (0.050, 0.03, 1.01), respectively, for case (ii).

Finally, simulations were conducted under a contiguous alternative hypothesis. All variables were simulated 1000 times exactly as for the setting in Table 1(a) except for the survival time, which was simulated as $\log X = (1 - R)\beta/n^{\frac{1}{2}} - W0.75 + \varepsilon$. Several choices of β were considered. For example, at the true values $\rho(R)$ and $\phi(x; R)$, for $n = 200$, with $\beta = 5.6$, the empirical powers of $L_n(\rho)$ and $\tilde L_n(\rho,\phi)$ were 0.64 and 0.89, respectively; with $\beta = 7.1$, their empirical powers were 0.86 and 0.98, respectively. Although it is in general more complex to conduct sensitivity analyses of the test $\tilde L_n(\rho,\phi)$, it can be more efficient than $L_n(\rho)$.

4. Example

Between January 1996 and January 1997, ACTG 320 enrolled patients to receive either the drug combination ZDV+3TC+placebo, with $R = 0$, or ZDV+3TC+indinavir, with $R = 1$ (Hammer et al., 1997). We analyze the ACTG 320 data as of February 18, 1997, the date of the interim analysis at which it was decided to stop accrual and close the study because of a significant beneficial effect of indinavir on the endpoint defined as the time to death or AIDS, whichever comes first. As previously mentioned, this analysis assumed that censoring was noninformative. The data used in this article differs slightly from that used in Hammer et al. (1997) because of retrospective updating; here, 581 patients were randomized to arm $R = 0$ and 575 patients were randomized to arm $R = 1$.

For groups $R = 0$ and $R = 1$, respectively, there were 19 and 11 deaths, and 57 and 29 AIDS events. For the event AIDS or death, whichever comes first, 66 occurred in $R = 0$ and 38 in $R = 1$. For each patient, the random variable C is defined as the calendar date February 18, 1997, minus the calendar date of enrollment; the median of C was 293 days for both treatment groups. The number of patients who were lost to follow-up before their respective administrative censoring time was 55 in the placebo group and 39 in the indinavir group. These numbers do not include deaths, since death was part of the definition of the primary event. In this case, for those patients observed to die, C^* is only known to lie between the time to death and C.

It is well known that the number of HIV-RNA copies per mm^3 of plasma, so-called viral load, is negatively associated with time to AIDS and death. The median baseline viral load was approximately 5 $\log_{10}$ units in both treatment groups. Of the 55 patients observed to prematurely discontinue follow-up in the placebo group, for those with baseline viral load below 5 the median follow-up time was 204 days and for those above 5 was 201 days. On the other hand, of the 39 patients in the indinavir group who prematurely discontinued follow-up, for those with baseline viral load below 5 the median follow-up time was 224 days and for those above 5 was 211 days. Thus, it may be the case that dropout depends on both treatment group and survival time.

For comparing treatment groups with respect to the time to AIDS or death, $L_n(1) = -2.8$ and $\tilde L_n(1,1) = -2.9$. Figure 1 displays a sensitivity analysis of $L_n(.)$ over $[0.56, 1] \times [0.51, 1]$; this range for $\widehat{\mathrm{pr}}(\delta = 1 \mid \delta^C = 1, R)$ was chosen because the smallest possible sample-mean estimates of these probabilities given the observed data valued 0.56 and 0.51 for $R = 0$ and $R = 1$, respectively. Here we have chosen to display the sensitivity analysis as a function of $\widehat{\mathrm{pr}}(\delta = 0 \mid \delta^C = 1, R)$ because, as stated in Section 1.2, we feel that this quantity may be easier to interpret. If it is plausible to assume that the probability that neither AIDS nor death is observed among those for whom one of these events occurs before analysis time is smaller for group $R = 1$ than for group $R = 0$, then the null hypothesis would be rejected.

5. Discussion

When the primary event is disease only, for example in studies where there is a negligible risk of death from relevant causes, the censoring time, $C^* = \min(C,\, D)$, is observed in full for each subject. In this case, Lin, Robins, and Wei (1996) proposed methods to test the effect of treatment group on the distribution of survival time, after adjusting for dependent censoring by assuming a bivariate location-shift model for the joint distribution of survival and censoring times.

The testing methodology proposed can be conducted separately for a small number of strata defined by baseline covariates. Incorporation of high-dimensional covariates into the testing methodology would require unverifiable modeling assumptions regarding the joint effect of treatment and covariates on the distribution of δ, and thus is not advocated in this article.

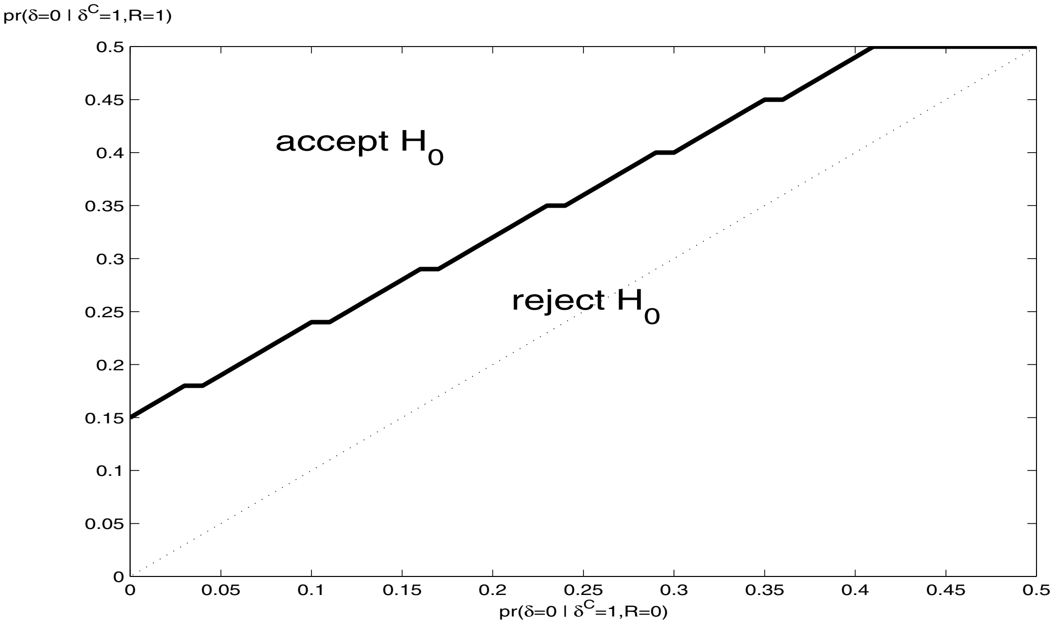

Figure 1. Test statistic $L_n(\cdot)$ as a function of possible sample-mean estimates of $\text{pr}(\delta = 0 \mid \delta^C = 1, R)$; above heavy line H_0 is not rejected, below heavy line H_0 is rejected.

The setting considered in this article is easily extended to the case when two randomized treatment groups are to be compared with respect to a possibly censored longitudinal outcome variable, measured at study time τ. For this setting, denote the positive longitudinal outcome variable at time τ by X_i; the random variable C_i is assumed to be independent of R_i and in this setting can be, for the example when X_i is a biomarker variable, the lower limit of quantification of X_i. Again, X_i^* denotes the observed portion of X_i, and $\delta^C = 1$ if $X_i^* = X_i$ and $\delta^C = 0$ otherwise. The nonresponse indicator Δ_i assumes the value $\Delta_i = 0$ if (X_i^*, δ^C) is missing for subject i and $\Delta_i = 1$ otherwise; also $\delta_i = \Delta_i \delta_i^C$. The observed data consists of the n independent and identically distributed realizations $(R_i, W_i, C_i, \Delta_i, \Delta_i X_i^*, \delta_i)$, $i = 1, \ldots, n$. Now, with the redefinitions $\rho(R_i) = 1/\text{pr}(\Delta = 1 \mid \delta^C = 1, R)$, $Y_i(x) = I(\Delta_i = 1, x \le X_i^*)$, $N_i(x) = I(\Delta_i = 1, x \ge X_i^*, \delta_i = 1)$, $\alpha(x) = \{\text{pr}(\Delta = 1 \mid X^* \ge x, R = 1)\}/\{\text{pr}(\Delta = 1 \mid X^* \ge x, R = 0)\}$, $i = 1, \ldots, n$, the tests $L_n(.)$ and $\tilde{L}_n(.,.)$ directly apply.

Note that when there is no censoring, the test $L_n(.)$ cannot be used for testing H_0.

Acknowledgements

This work was supported in part from grants R01AI28076 and R01AI51164-01 from the U.S. National Institutes of Health. The author is grateful to an associate editor and two referees, whose comments greatly improved the presentation of the article.

Résumé

Quand on teste l'hypothèse nulle que les distributions des durées de survie spécifiques de groupes de traitements sont égales, le test du logrank est asymptotiquement valide quand la distribution du temps de censure est conditionnellement indépendante du groupe de traitement randomisé sachant le temps de survie. Nous introduisons un test d'hypothèse nulle utilisable lorsque la distribution des temps de censure dépend du groupe de traitement et du temps de survie. Ce test ne fait aucune supposition concernant l'indépendance du temps de censure et de la durée de survie. La validité asymptotique de ce test requiert seulement un estimateur consistant pour la probabilité conditionnelle que l'événement soit observé, sachant à la fois le groupe de traitement et le fait que l'événement se soit produit avant le temps de l'analyse. Cependant, si l'on ne fait pas de supposition (invérifiable) concernant le mécanisme de génération des données pour chaque groupe de traitement, il existe un ensemble de valeurs possibles pour les estimations correspondantes de la moyenne de ces probabilités qui sont compatibles avec les données observées. Sur ce sous-ensemble du carré unité, le test proposé peut être calculé, et une région de rejet identifiée. Une décision concernant l'hypothèse nulle, qui prend en compte l'incertitude due au fait que la censure peut dépendre du groupe de traitement et de la durée de survie, peut être prise directement. Nous présentons également un test du logrank généralisé qui nous permet de fournir les conditions sous lesquelles le test du logrank ordinaire est asymptotiquement valide. Le test généralisé peut être aussi utilisé pour tester l'hypothèse nulle quand la distribution de la censure dépend du groupe de traitement et de la durée de survie. Cependant l'usage de ce test nécessite des suppositions de modélisation semi-paramétrique. Une étude par simulation, et un exemple utilisant un essai clinique récent sur le SIDA sont fournis.

References

DiRienzo, A. G. and Lagakos, S. W. (2001a). Bias correction for score tests arising from misspecified proportional hazards regression models. *Biometrika* **88,** 421–434.

DiRienzo, A. G. and Lagakos, S. W. (2001b). Effects of model misspecification on tests of no randomized treatment effect arising from Cox's proportional hazards model. *Journal of the Royal Statistical Society, Series B* **63,** 745–757.

Fisher, L. and Kanarek, P. (1974). Presenting censored survival data when censoring and survival times may not be independent. In *Reliability and Biometry: Statistical Analysis of Lifelength,* F. Proschan and R. Serfling (eds), 303–326. Philadelphia: SIAM.

Hammer, S. M., Squires, K. E., Hughes, M. D., et al (1997). A controlled trial of two nucleoside analogues plus indinavir in persons with HIV infection and CD4 cell counts of 200 per cubic millimeter or less. *New England Journal of Medicine* **337,** 725–733.

Klein, J. P. and Moeschberger, M. L. (1988). Bounds on net survival probabilities for dependent competing risks. *Biometrics* **44,** 529–538.

Klein, J. P., Moeschberger, M. L., Li, Y. H., and Wang, S. T. (1992). Estimating random effects in the Framingham Heart Study (with discussion). In *Survival Analysis: State of the Art,* J. Klein and P. Goel (eds), 99–120. Dordrecht: Kluwer.

Lin, D. Y., Robins, J. M., and Wei, L. J. (1996). Comparing two failure time distributions in the presence of dependent censoring. *Biometrika* **83,** 381–393.

Moeschberger, M. L. and Klein, J. P. (1995). Statistical methods for dependent competing risks. *Lifetime Data Analysis* **1,** 195–204.

Robins, J. M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proceedings of the Biopharmaceutical Section, American Statistical Association,* 24–33. Alexandria, Virginia: American Statistical Association.

Robins, J. M. and Finkelstein, D. H. (2000). Correcting for non-compliance and dependent censoring in an AIDS clinical trial with inverse probability of censoring weighted (IPCW) log-rank tests. *Biometrics* **56,** 779–788.

Robins, J. M. and Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In *AIDS Epidemiology: Methodological Issues,* N. Jewell and K. Dietz (eds), 297–331. Boston: Birkhauser.

Satten, G. A., Datta, S., and Robins, J. M. (2001). An estimator for the survival function when data are subject to dependent censoring. *Statistics and Probability Letters* **54,** 397–403.

Scharfstein, D. O. and Robins, J. M. (2002). Estimation of the failure time distribution in the presence of informative censoring. *Biometrika* **89,** 617–634.

Scharfstein, D. O., Rotnitzky, A., and Robins, J. M. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association* **94,** 1096–1146.

Slud, E. V. and Rubinstein, L. V. (1983). Dependent competing risks and summary survival curves. *Biometrika* **70,** 643–649.

Zheng, M. and Klein, J. P. (1995). Estimates of marginal survival for dependent competing risks based on an assumed copula. *Biometrika* **82,** 127–138.

Received March 2002. *Revised February* 2003.
Accepted February 2003.