Published in final edited form as:

J Stat Plan Inference. 2008 July 1; 138(6): 1722–1732. doi:10.1016/j.jspi.2007.06.028.

Testing Equality of Survival Distributions when the Population Marks are Missing

Dipankar Bandyopadhyay^{a,*} and Somnath Datta^b

^a Department of Biostatistics, Bioinformatics and Epidemiology, Medical University of South Carolina, Charleston, SC 29425

^b Department of Bioinformatics and Biostatistics, University of Louisville, Louisville, KY 40292

Abstract

This paper introduces a nonparametric approach for testing the equality of two or more survival distributions based on right censored failure times with missing population marks for the censored observations. The standard log-rank test is not applicable here because the population membership information is not available for the right censored individuals. We propose to use the imputed population marks for the censored observations leading to fractional at-risk sets that can be used in a two sample censored data log-rank test. We demonstrate with a simple example that there could be a gain in power by imputing population marks (the proposed method) for the right censored individuals compared to simply removing them (which also would maintain the right size). Performance of the imputed log-rank tests obtained this way is studied through simulation. We also obtain an asymptotic linear representation of our test statistic. Our testing methodology is illustrated using a real data set.

Keywords

Equality of survival curves; Fractional at-risk set; Log-rank tests; Multistate models

1. Introduction

Testing equality of two or more survival curves is a well studied problem in statistics. The resulting methodologies, e.g., log-rank tests and Cox's regression constitute some of the frequently used statistical techniques while dealing with survival or, more generally, time to event data. The setup assumes that one has independent samples from two or more populations whose survival functions are to be compared. Of course, most of the 'time to event' data are subject to some form of censoring, right censoring being the most common one. A unified theoretical and methodological treatment of these problems is achieved through a counting process and related martingale techniques.

One basic assumption in using the log-rank type test is that every individual is clearly identified to have come from one of two or more independent populations including those whose failure times are right censored. Now consider a situation where the population marks are also unknown, along with their failure times, for the right censored individuals. For example, consider a multistate model setup (see page 26 in Andersen et al., 1993) where a healthy (state 0) subject may end up in one of the J stages corresponding to different types of death. If we are interested in studying the sub-populations corresponding to the different death types, often,

*Corresponding author. Tel.: +1 843 876 1603, Fax: +1 843 876 1126, bandyopd@musc.edu.

the membership information can only be found after the individual actually dies (e.g., through autopsy, say). As a result, these important sub-population marks will be unavailable for individuals who were right censored during an observational study. As for example, Crowley and Hu (1977) studies patients from the popular Stanford Heart Transplantation Program where transplant recipients were classified into subpopulations based on the cause of failure (here death), viz., transplant rejection, or other causes. A pertinent question to address is whether the survival curves for the two different subpopulations are the same, or not. We revisit this data set and answer this question in Section 4.

Another example where a group or population mark is not known in the beginning of a study may arise in a toxicological experiment as follows. Consider an experiment on animals where one is interested in studying the effect of a certain carcinogenic agent which can also produce other adverse effects leading to death of an animal before it actually causes cancer. Amongst other things, the survival distribution of animals dying from cancer caused by this agent was of interest. Animals are autopsied at death to study the tumor growth and type through which their population mark was determined where the sub-population of interest is that of all animals where the toxicological agent would cause a cancer death. Suppose, for some reason, the necessary autopsy could not be performed on certain animals although their death times were known. Our approach would apply to this situation where we would assign a fractional mass to each such individual towards the at-risk set of cancer death. Next suppose, one is comparing the survival times of animals dying of cancer caused by two such agents applied to two sets of animals. Suppose, once again, for certain animals in each set, we fail to determine whether they died of cancer or some other adverse effect. Our testing methodology would handle such individual by assigning them fractional masses. Another approach would be to simply remove the censored observations and compare the survival times of the animals that are known to have died of cancer. However, as shown by a simple example in the next section, removing censored observations may result in a loss of power.

There are only a limited number of papers dealing with the testing problem we are interested in. Related papers by Dykstra et al. (1998) and Kocher and Proschan (1991) provide some restricted tests for testing the independence of time to failure and cause of failure in a competing risk framework. Testing of dependence structures between failure time and cause of failure expressed in terms of monotonicity properties of the conditional probabilities involving failure times and failure cause are proposed very recently by Dewan et al. (2004). They used a U -statistic approach which assumes no censored observations.

The rest of the article is organized as follows. In Section 2, we develop our testing methodology. Section 3 deals with two simulation studies to check the finite sample performance of our test statistic under different alternatives. Section 4 illustrates applications of our proposed test to a real life data set mentioned before. The paper ends with a discussion section (Section 5) followed by an appendix containing an outline of an asymptotic linear representation of our test statistic.

2. Tests using Fractional Risk Sets

2.1. Notation

We consider a multistate continuous time stochastic process $\{Z(t), t \in \mathcal{T}\}$ with a finite state space $\mathcal{S} = \{1, \dots, J, 0\}$ having a tree topology and right-continuous sample paths: $Z(t+) = Z(t)$ where we assume that the states $1, \dots, J$ are absorbing whereas state 0 is transient (the root node). Here $\mathcal{T} = [0, \tau]$ where τ is a large possibly observed time point ($\leq \infty$). Typically, for applications, τ will be taken to be the largest time where some event (failure) took place. Let T_i^* be the (possibly unobserved) time the i th person leaves state 0 for a failure (state j , say). Let X_i^* denote the state occupied by the i th individual at time T_i^* (i.e., its failure type). Let C_i be the

censoring time for the i th person. Let $T_i = \min(T_i^*, C_i)$ denote the right censored failure time and δ_i denote the failure/censoring indicator

$$\delta_i = \begin{cases} j & \text{if } T_i^* \leq C_i \text{ and } X_i^* = j, \\ 0 & \text{if } T_i^* > C_i. \end{cases} \quad (1)$$

Note that δ_i is an observable quantity for everybody but X_i^* is observed only for the uncensored people for whom the two are equal. It is further assumed that the censoring variable C_i is independent of the entire collection of $\{T_i^*, X_i^*\}$ and all the random variables are i.i.d. across the n individuals. Our data consist of the pairs (T_i, δ_i) , $1 \leq i \leq n$.

Define the conditional survival function $S_j(t)$ as

$$S_j(t) = P(T_i^* > t | X_i^* = j), \quad 1 \leq j \leq J.$$

We would like to construct an omnibus test for the null hypothesis,

$$H_0: S_1(t) = \dots = S_J(t) (\equiv S(t), \text{ say}), \quad \forall 0 \leq t \leq L,$$

where $L \leq \tau$ is a finite but large time point.

2.2. The Test Statistic

A standard log-rank test would have been applicable if the individuals could be separated into various subpopulations depending on their failure types. Basically, the Nelson-Aalen estimators of the (conditional) cumulative hazards for various sub-populations could be compared amongst each other. However, as indicated earlier, the difficulty now is that the sub-population memberships X_i^* 's are not known for the censored people and simply including them in at-risk consideration for all sub-populations will lead to a biased comparison leading to an improper size for the resulting test in case of unbalanced designs (see section 2.5).

Since a log-rank test basically compares the estimated hazards of various sub-populations, it will be helpful to recall the definition of Nelson-Aalen estimator (Andersen et al., 1993) of cumulative hazards of failure amongst individuals of sub-population (or failure type) j . Let N_j be the counting process counting the number of observed failures of type j (i.e., in sub-population j) in the time interval $[1, t]$:

$$N_j(t) = \sum_{i=1}^n I(T_i \leq t, \delta_i > 0, X_i^* = j),$$

and let $Y_j(t)$ denote the number of individuals in the j th sub-population that are at-risk of failing or getting censored at time t :

$$Y_j(t) = \sum_{i=1}^n I(T_i \geq t, X_i^* = j).$$

As a stochastic process, Y_j is predictable, i.e., $Y_j(t)$ is $\mathcal{F}_{t-}$ measurable, where $\mathcal{F}_t = \sigma([N_j(s), \sum I(T_i \leq s, \delta_i = 0, X_i^* = j) : 0 \leq s \leq t, j \geq 1])$. Then the Nelson-Aalen estimator for the j th sub-population is given by

$$\widehat{\Lambda}_j(t) = \int_0^t \frac{I(Y_j(s) > 0)}{Y_j(s)} dN_j(s). \quad (2)$$

It is not difficult to see that N_j is computable from the observed data, since $\{\delta_i > 0, X_i^* = j\}$ is the same as $\{\delta_i = j\}$. The size of the sub-population at-risk set $Y_j(t)$ on the other hand, is not computable, once again, due to the unavailability of all the sub-population marks. Thus, we estimate $Y_j(t)$ by $Y_j^f(t)$, where $Y_j^f(t)$ is the size of a ‘fractional’ or ‘imputed’ at-risk set corresponding to the j th sub-population defined as (Satten and Datta, 1999; Datta et al., 2000)

$$Y_j^f(t) = \sum_{i=1}^n \widehat{\varphi}_{ij} I(T_i \geq t), \quad (3)$$

where $\widehat{\varphi}_{ij}$ is the estimated probability that the i th individual belongs to the j th sub-population based on the available data. It is not hard to see that, under independent censoring, a reasonable choice for $\widehat{\varphi}_{ij}$ is given by

$$\widehat{\varphi}_{ij} = \begin{cases} \widehat{P}_j(T_i, \infty), & \text{if } \delta_i = 0 \\ 1, & \text{if } \delta_i = j \\ 0, & \text{if } \delta_i > 0, \delta_i \neq j, \end{cases} \quad (4)$$

where $\widehat{P}_j(T_i, \infty)$ is the nonparametric maximum likelihood estimator (NPMLE) or the Aalen-Johansen estimator of the transition probability

$$P_j(s, t) = \Pr\{T_i^* \leq t, X_i^* = j | T_i^* > s\} \text{ at } s = T_i, t = \infty.$$

Specializing Andersen et al. (1993) results for a Markov chain to a competing risk setup, we obtain

$$\widehat{P}_j(s, t) = \int_{(s, t]} \left\{ \prod_{(s, u)} \left(1 - \frac{dN_j(v)}{Y(v)} \right) \right\} \frac{dN_j(u)}{Y(u)}. \quad (5)$$

In the expression (5), $Y(t) = \sum_{i=1}^n I(T_i \geq t)$ is the size of the at-risk set irrespective of failure types

and $N(t) = \sum_{j=1}^J N_j(t)$ is the total number of observed failures of all types (total number of states

entered) by time t . Since N_j and N are discrete with jumps only at the failure times T_i 's, the above integral can be replaced by sums leading to the following simpler expression

$$\widehat{P}_j(s, t) = \sum_{s < T_i \leq t} \left\{ \frac{\widehat{S}(T_i^-)}{\widehat{S}(s)} \right\} \left\{ \frac{\Delta N_j(T_i)}{Y(T_i)} \right\},$$

where $\widehat{S}$ is the Kaplan-Meier estimator of failure due to all causes and $\Delta N_j(T_i)$ is the number of failures of type j at time T_i . Note that for a typical individual who had not failed up to and including time s , the first term $\widehat{S}(T_i^-)/\widehat{S}(s)$ in the above summand computes the probability that such an individual had not failed until time T_i and the second term $\Delta N_j(T_i)/Y(T_i)$ is probability of its failing at time T_i and the failure is of type j given survival until that time. See Datta and Satten (2000) for theoretical properties of the above estimator (4), which would show that,

asymptotically, $Y_j^f(t)$ is a good approximation to $Y_j(t)$; e.g., the ratio $Y_j^f(t)/Y_j(t) \xrightarrow{P} 1$.

A natural class of a nonparametric test process $\{\Delta_j(t), t \geq 0\}$ that are applicable to such a setup is the class of weighted log-rank tests based on the weighted departures of the estimated Λ_j compared to a pooled estimate of the common Λ under the null hypothesis given by

$$\Delta_j(L) = \int_0^L W(t) \left\{ dN_j(t) - \frac{Y_j^f(t)}{Y^f(t)} dN(t) \right\} \quad (6)$$

where $W(t)$ is a non-negative, locally bounded, predictable weight process that provides a flexible way to control the relative importance attached to differences in the estimated

conditional hazards at various time points, $Y_j^f(t)$ is as defined in (3) and $Y^f(t) = \sum_{j=1}^J Y_j^f(t)$ be the total fractional risk set at time t .

The components of the vector $\underline{\Delta} = (\Delta_1(L), \dots, \Delta_J(L))$ are linearly dependent since

$\sum_{j=1}^J \Delta_j(L) = 0$. Let $\widehat{\Sigma}$ be an estimated variance covariance matrix of the vector of test statistic $\sqrt{n}\underline{\Delta}$ and let $\widehat{\Sigma}^-$ be its spectral (generalized) inverse. Our test statistic is given by the quadratic form

$$T = n \underline{\Delta}^T \widehat{\Sigma}^- \underline{\Delta} \quad (7)$$

which has an asymptotic chi-squared distribution with $J - 1$ degrees of freedom under the null hypothesis provided suitable regularity conditions are met.

2.3. Asymptotic Normality and Variance Estimation

The most commonly used tool for obtaining asymptotic distributions of estimators and test statistics with right censored data is the counting process formulation and the associated martingale techniques pioneered by Odd Aalen (Aalen, 1978) and used by numerous authors since then. Certainly they work for the regular log-rank tests when subpopulation marks are

known. However, in the present context, the imputed subpopulation marks are based on the entire data set and in a sense the fractional at-risk set at time t are functions of future information as well (i.e., they are not predictable with respect to the natural filtration process one considers). However we can still obtain asymptotic normality of the test statistic by establishing the following asymptotic linear representation:

$$\Delta_j(L) = \sum_{i=1}^n V_{j,i} + o_p(\sqrt{n}), \quad (8)$$

where

$$\begin{aligned} V_{j,i} = & \int_0^L w(t) dM_{j,i}(t) - \int_0^L w(t) \frac{y_{j,i}(t)}{y(t)} dM_{\cdot,i}(t) + \\ & \int_0^L \frac{w(t)}{y(t)} \left\{ I(T_i \geq t, X_i = j, \delta_i = 0) - P_j(T_i, \infty) I(T_i \geq t, \delta_i = 0) \right. \\ & + \int_t^\infty \int_z^\infty \frac{S(u)}{S(z)y(u)} P_j(u, \infty) dM_{\cdot,i}(u) dn^C(z) \Big\} dn(t) \\ & - \int_0^L \frac{w(t)y_{j,i}(t)}{y^2(t)} \sum_{j'=1}^J \left\{ I(T_i \geq t, X_i = j', \delta_i = 0) - P_{j'}(T_i, \infty) I(T_i \geq t, \delta_i = 0) \right. \\ & + \left. \int_t^\infty \int_z^\infty \frac{S(u)}{S(z)y(u)} P_{j'}(u, \infty) dM_{\cdot,i}(u) dn^C(z) \Big\} dn(t), \end{aligned}$$

with

$$\begin{aligned} M_{j,i} = & I(T_i \leq t, \delta_i > 0, X_i = j) - \int_0^t I(T_i \geq s, X_i = j) d\Lambda(s), M_{\cdot,i} = \sum_j M_{j,i}, \\ y_{j,i}(t) = & P(T_i \geq t, X_i = j); y(t) = \sum_j y_{j,i}(t), \\ n(t) = & P(T_i \leq t, \delta_i > 0), n^C(t) = P(T_i \leq t, \delta_i = 0), \end{aligned}$$

and w is the (in-probability) limit of the weight function W .

We present an outline of the derivation of (8) in the appendix. This representation also yields a closed form expression for $\hat{\Sigma}$ obtained by the empirical variance-covariance of $\hat{V}_{j,i}$, where $\hat{V}_{j,i}$ is obtained by replacing all the population quantities in $V_{j,i}$ by their estimates. However, a close examination of the formulas reveal that the resulting expression is computationally cumbersome and requires $O(n^3)$ order of calculations. A more practical solution will be to use a resampling (bootstrap) scheme to compute the estimated variance-covariance matrix $\hat{\Sigma}$. Asymptotic validity (consistency) of this bootstrap based estimate can be established by deriving a similar linear representation (8) for the resampled statistic. The theoretical details are not presented here.

2.4. Variance Estimation by Bootstrap

For testing problems, it is preferred that the resampling scheme incorporates the null hypothesis. This is achieved as follows:

- i. Let $\delta'_i = I(\delta_i > 0)$ denote the true failure indicators irrespective of the failure types, $1 \leq i \leq n$. Generate i.i.d. resamples $(\hat{T}_i^*, \hat{\delta}_i^{*'})$, $1 \leq i \leq n$, from the empirical distribution of the pairs $\{(T_i, \delta'_i) : 1 \leq i \leq n\}$.

- ii. We let $\widehat{\delta}_i^* = 0$, if $\widehat{\delta}_i^{*'} = 0$; otherwise, for $\widehat{\delta}_i^{*'} > 0$, generate $\widehat{\delta}_i^*$ from a J -point distribution with $\text{Prob}\{\widehat{\delta}_i^* = j\} = \widehat{\Phi}_j / \sum_{k=1}^J \widehat{\Phi}_k$, with

$$\widehat{\Phi}_j = \sum_{i=1}^n \widehat{\varphi}_{ij}, 1 \leq j \leq J.$$

A typical bootstrap sample is given by $(\widehat{T}_i^*, \widehat{\delta}_i^*), i = 1, \dots, n$.

- iii. Repeat steps (i) and (ii), a large number of times, say B , to obtain B sets of bootstrap samples $(\widehat{T}_{i,b}^*, \widehat{\delta}_{i,b}^*)$ and compute the corresponding test statistics $\widehat{\Delta}_b^*$, $b = 1, \dots, B$.

A bootstrap estimate of $\widehat{\Sigma}$ is given by the empirical variance-covariance matrix of the $\widehat{\Delta}_b^*$, $1 \leq b \leq B$. In the next section, we perform a number of simulation studies to assess the performance of our test under two alternative hypotheses models.

2.5. Reasons for Imputing Marks

We now present two simple examples to illustrate the reasons for imputing the population marks for the censored observations leading to the fractional masses in calculating the at-risk sets for the two populations.

The first example considers the option of removing all censored observations in comparing the populations. This option would maintain the correct size in the resulting log-rank test but may lead to a loss in power. We generated equal number of observations from two populations. Failure times in Population 1 were generated from a Uniform(0.9, 1.2) distribution. For the size comparison, we generated the failure times in Population 2 from the same distribution and for the power comparison we generate them from Uniform(0.2 + i , 3.1) distribution where $i = 0(0.1)0.3$. The four values of i create four alternatives for the testing problem that we are interested in. Censoring times were independently generated from a Uniform (0.9, 3.0) distribution. We choose a sample size of $n = 50$, a targeted nominal level of $\alpha = 0.05$, a bootstrap size of $B = 1000$ and a Monte Carlo iteration size of $N = 2000$ to compute powers empirically. Table 1 reports the power of our log-rank test based on fractional masses for censored observations and that of a log-rank test where censored observations are removed from the sample. We observe that while both tests achieve the nominal 0.05 level, there is nearly 30% power gain by imputing their population marks rather than removing the censored observations.

Next we illustrate what happens if we ignore the population marks of the censored observations and they each count a whole mass in the at-risk sets for both populations, $Y_j, j = 1, 2$. This will indeed be the case if one uses the standard competing risk methodology which tests the equality of cause specific hazard rates where the cause of death is the population mark that is revealed at death. We simulated $n = 100$ lifetimes from a standard exponential distribution. For each lifetime, a population mark of 1 or 2 is randomly assigned with probability ϕ of being in population 1. Note that, by construction, the survival times in both population follow the same distribution (standard exponential) in this case. Censoring times were 2.4 times standard exponentials leading to a 30% censoring rate. Size of a 5% log-rank test where censored observations are counted in both at-risk set are computed empirically using $N = 2000$ Monte Carlo runs each. The following table clearly demonstrates that this approach would result in an incorrect size.

As illustrated by the second example, it is important to think about precisely what null hypothesis is to be tested. If we are indeed interested in testing the equality of the two survival functions then treating this as a competing risk problem may incorrectly reject the null hypothesis simply because the data is unbalanced.

3. Simulation Studies

To assess the performance of our test statistic, we conducted two different simulation schemes to be described next each resulting in a multistate model with $J = 2$ terminal states corresponding to two sub-populations. We used sample sizes of $n = 50$ and 200 . A nominal level $\alpha = 0.05$ was used in each case. Furthermore, for each sample, a bootstrap replication size of $B = 1000$ was used. Finally the power was computed by the proportion of rejection out of $N = 2000$ Monte-Carlo replications. The weight function was taken to be unity and the largest observed failure time was used for L in all cases.

3.1. Simulation Schemes

(i) Fixed sub-population proportions—Under this scheme, we generate data in such a way that the relative proportion of the two sub-populations remain constant no matter what sample size is chosen. This is done to illustrate the effect of sample size on power while other settings remain comparable. We first simulate the sub-population indicators X_i^* by setting the arm probability φ to be 0.25, 0.5 or 0.75 and letting X^* to be 1 if a randomly generated uniform (0,1) random variable falls below φ , and 2 otherwise. Next, lifetimes T_i^* were generated from a standard lognormal ($= \exp(N(0,1))$) distribution for both sub-populations under the null hypothesis. For the alternative hypothesis, lifetimes for the sub-population $X^* = 1$ were generated as in the null case; however, samples for the $X^* = 2$ sub-populations were generated as $\exp(N(-a, 1))$ with $a = 0(0.3)1.5$, where the value in parentheses indicate increments. The censoring times C_i were generated from the lognormal distributions with the mean parameters 0.95 and 0, while keeping the variance parameter to be 1 in both cases, representing light (about 25%) and heavy (about 50%) censoring, respectively.

(ii) Random subpopulation proportions—For this scheme we consider an individual being exposed to two dependent risks and let the notional (or latent) lifetimes of the unit under the two risks be denoted by L_1 and L_2 . The lognormal distribution was chosen to generate L_1 and L_2 in order to concretely model dependence through the bivariate normal distribution. These result in failure time data and population marks given by $T^* = \min(L_1, L_2)$ and $X^* = 2 - I(L_1 \leq L_2)$. We let $L_1 = \exp(L'_1)$ and $L_2 = \exp(L'_2)$, where (L'_1, L'_2) are generated from a bivariate normal distribution as follows. For the null hypothesis, we let the mean of this bivariate normal distribution be (0,0), and variance-covariance matrix be $((1, \rho), (\rho, 1))$; three choices of ρ , e.g., $-0.5, 0$ or 0.5 , are considered. For the alternative, we generate (L'_1, L'_2) from a bivariate normal distribution with mean vector $(0, (1 - a))$ and variance-covariance matrix $((1, \rho a), (\rho a, a^2))$ where $a = 1(0.3)2.5$. The censoring times were generated from lognormal distributions with variance parameter 1 and mean parameters 0.95 and 0 leading to light (8%–17%) and moderately heavy (28%–45%) censoring, respectively.

3.2. Results

First, we report the empirical sizes of our imputed log-rank tests for the two simulation schemes (Tables 3 and 4). The empirical sizes of our tests are within two standard errors of the nominal size for the larger sample size ($n = 200$). The empirical sizes for $n = 50$ are also fairly close to the nominal level although they are marginally inflated in a few cases.

Figure 1 display an array of plots illustrating the power curves under scheme (i) as a function of the alternative parameter a which was defined before. Once again, the set up was identical

as the size study above and includes two choice of sample sizes, namely, 50 and 200. We overlay the power curves obtained by our method with the power curves of a regular log-rank test if the sub-population information were known for all individuals. It is seen that a considerable increase in power is observed with an increase in sample size. The power curves reveal that for low censoring and larger sample size, our imputed test produces a power profile that is very close to the log-rank test with the additional information about population marks for the censored observation. Heavy censoring has some adverse effect on the empirical powers which tend to get better with the larger sample size. Power curves from scheme (ii) reveal similar conclusions and are not presented here.

As mentioned earlier, all the values reported in Tables 3 and 4 were based on $B = 1000$, which was judged sufficient after balancing accuracy with computational burden. We have studied the sensitivity of our results with respect to the selection of the bootstrap replication size B . Power calculations for the imputed log-rank test using $B = 500$ turned out to be very close to the tables presented here; we omit the details.

4. An Example

We now illustrate our test using the Stanford Heart Transplant data mentioned in the introduction.

We consider the same subset of $N = 65$ patients of the Stanford Heart Transplant data (Crowley and Hu, 1977) as considered in Larson and Dinse (1985). The study began in October of 1967 and patients were admitted to the Stanford program for heart replacement. Among these 65 patients who received transplants, there were 29 (45%) rejection deaths (RD), 12 (18%) deaths from other causes (OC), and 24 (37%) right censored observations. The value of τ was 1775 days and the minimum event time was 0 days. The survival times of RD patients ranged from 10 days to 1350 days and that for the OC patients from 0 to 551 days.

We used $B = 5000$ bootstrap replicates to compute the standard error of the test statistic. The value of Δ for this data set was 7.07 with a bootstrap estimated standard error of 3.58 leading to a χ^2 test statistic value of 3.92 that was significant at 5%. Figure 2 shows a plot of the cumulative hazards (Nelson-Aalen) of the two disease types using fractional risk sets. We can conclude that, in this population of patients, those who die of other causes tend to have a shorter life span than those who die of transplant rejection.

5. Discussion

A weighted log-rank test for testing the equality of J survival functions has been proposed when the population marks of the right censored individuals in the study are unknown. Such a structure arises naturally in practice. The concept of fractional or imputed membership was the key in breaking up the total at risk set into J sub-at-risk sets (corresponding to J (sub) populations) at every event time point; that's how we connect this testing problem to the traditional log-rank testing methodology.

Due to the lack of a suitable martingale representation, the form of the estimated variance-covariance matrix of the test statistics is computationally cumbersome and as a result we propose a bootstrap scheme to carry out the test by computing the bootstrap variance-covariance matrix. Our test is seen to achieve satisfactory performance as compared to a log-rank test where the population marks are known for the censored individuals in terms of size and power for small to moderate sample size.

It is easy to see that, if one chooses to ignore the sub-population information for censored data and let them contribute full mass (whole observations) to all at-risk set then the resulting test

could have wrong size for our null hypothesis. This will be true whenever the sub-population are of unequal percentages of the whole population. If on the other hand, the censored observations are thrown away (equivalent to contributing zero mass to each at-risk set), the size would be maintained but there may be a loss in power. This was illustrated in Table 1.

Although in this paper we restrict our attention to the class of log-rank tests, the concept of fractional at risk set goes much further. In essence, it will enable us to propose versions of other tests for testing equality of survival curves based on independent right-censored samples that are based on at risk sets, where the population mark is known for everyone, to the present situation. Examples of such tests would include tests based on weighted difference of Kaplan-Meier survival estimates of conditional survival functions for the J subpopulations. The conditional survival functions for the J sub-populations can be obtained by normalizing the Aalen-Johansen estimators or equivalently by Kaplan-Meier formulas using the fractional risk sets. See Satten and Datta (1999) for this equivalence.

The present methodology does not consider any covariate. It might be of interest in some applications to ask whether the conditional survival functions in various sub-populations are equal given some available covariates. This will be pursued elsewhere.

Acknowledgments

The first author would like to acknowledge the research support from the University of Georgia for awarding him a Graduate School Dissertation Completion Fellowship. This research was supported in part by NIH/NCRR Grant P20 RR017696-04 and NSA Grant H98230-06-1-0062. We also thank an anonymous reviewer, the Associate Editor, and the Editor, for carefully reading the manuscript and for providing us with invaluable comments and suggestions which led to improvements.

References

- Aalen OO. Nonparametric inference for a family of counting processes. *Ann Statist* 1978;6:701–726.
- Andersen, PK.; Borgan, Ø.; Gill, RD.; Keiding, N. *Statistical Models Based on Counting Processes*. Springer-Verlag; New York: 1993.
- Crowley J, Hu M. Covariance analysis of heart transplant survival data. *J Amer Statist Assoc* 1977;72:27–36.
- Datta S, Satten GA. Estimating future stage entry and occupation probabilities in a multistage model based on randomly right-censored data. *Statist Probab Lett* 2000;50:89–95.
- Datta S, Satten GA, Datta S. Nonparametric estimation for three-stage irreversible illness-death model. *Biometrics* 2000;56:841–847. [PubMed: 10985224]
- Dewan I, Deshpande JV, Kulathinal SB. On testing dependence between time to failure and cause of failure via conditional probabilities. *Scand J Statist* 2004;31:79–91.
- Dykstra R, Kochar SC, Robertson T. Restricted tests for testing independence of time to failure and cause of failure in a competing risks model. *Canad J Statist* 1998;26:57–68.
- Kochar SC, Proschan F. Independence of time and cause of failure in the multiple dependent competing risks model. *Statist Sinica* 1991;1:295–299.
- Larson MG, Dinse GE. A mixture model for the regression analysis of the competing risk data. *Appl Statist* 1985;34:201–211.
- Satten GA, Datta S. Kaplan-Meier representation of competing risk estimates. *Statist Probab Lett* 1999;42:299–304.

Appendix A. An outline of the derivation of the linear representation

Using two ‘method of differences’ and the fact that the conditional hazard rates are the same under the null hypothesis we obtain from (6)

$$n^{-1/2}\Delta_j(L)=n^{-1/2}\int_0^L w(t)dM_j(t)-n^{-1/2}\int_0^L w(t)\frac{y_j(t)dM(t)}{y(t)}+n^{1/2}\int_0^L w(t)\left\{\frac{Y_j(t)}{Y(t)}-\frac{Y_j^f(t)}{Y^f(t)}\right\}dn(t)+o_p(1), \quad (\text{A.1})$$

under H_0 , where $dM_j(t) = dN_j(t) - Y_j(t)d\Lambda(t)$, $dM(t) = \sum_j dM_j(t)$, Λ denotes the common conditional cumulative hazard rate under the null hypothesis and $n(t)$, $y(t)$, and $y_j(t)$ are the in probability limit of the averages $n^{-1}N(t)$, $n^{-1}Y(t)$, and $n^{-1}Y_j(t)$, respectively.

By Corollary 3.1 in Datta and Satten (2000), we have $n^{-1}\{Y_j(t) - Y_j^f(t)\} \xrightarrow{P} 0$. Therefore, we can express the third term on the RHS of (A.1) as

$$n^{1/2}\int_0^L \left\{ \frac{w(t)}{y(t)}\{Y_j(t) - Y_j^f(t)\} - \frac{w(t)y_j(t)}{y^2(t)}\{Y(t) - Y^f(t)\} \right\} dn(t) + o_p(1). \quad (\text{A.2})$$

Next, by definition,

$$\begin{aligned} & n^{-1/2}\{Y_j(t) - Y_j^f(t)\} \\ &= n^{-1/2}\left\{ \sum_{i=1}^n I(T_i \geq t, X_i=j, \delta_i=0) - \sum_{i=1}^n \widehat{P}_j(T_i, \infty)I(T_i \geq t, \delta_i=0) \right\} \\ &= n^{-1/2}\sum_{i=1}^n \{I(T_i \geq t, X_i=j, \delta_i=0) - P_j(T_i, \infty)I(T_i \geq t, \delta_i=0)\} + \\ & \quad n^{-1/2}\sum_{i=1}^n \{P_j(T_i, \infty) - \widehat{P}_j(T_i, \infty)\}I(T_i \geq t, \delta_i=0) \end{aligned} \quad (\text{A.3})$$

Now express the second term on the RHS of (A.3) as

$$-n^{1/2}\int_0^\infty \{\widehat{P}_j(z, \infty) - P_j(z, \infty)\}I(z \geq t)dn^C(z) + o_p(1), \quad (\text{A.4})$$

The expression within the braces of (A.4) is the $(0, j)$ -th element of the matrix $\widehat{\mathbf{P}}(s, t) - \mathbf{P}(s, t)$ for $s = z$ and $t = \infty$. Therefore by Andersen et al. (1993, 4.4.6), we have after some algebra

$$\begin{aligned} & n^{1/2}\{\widehat{P}_j(z, \infty) - P_j(z, \infty)\} = \\ & -n^{1/2}\sum_{i=1}^n \int_z^\infty \frac{S(u)}{S(z)y(u)} P_j(u, \infty) dM_{\cdot, i}(u) + o_p(1). \end{aligned} \quad (\text{A.5})$$

Combining (A.3), (A.4), and (A.5), we get the linear representation for the first part of the expression with braces of RHS (A.2). A linear representation for the second part follows in the same way by summing over j' since $Y(t) - Y^f(t) = \sum_{j'} \{Y_{j'}(t) - Y_{j'}^f(t)\}$.

Figure 1.

Power curves of our tests in Simulation scheme (i). Upper two curves are for sample size 200 and lower two for sample size 50. The solid lines correspond to the standard log-rank tests with the knowledge of the population marks for everyone and the dashed lines correspond to our imputed log-rank test.

Figure 2. Estimated cumulative hazard rates for the two sub-populations in the Stanford Heart Transplant Data.

Table 1

Estimated power of our imputed log-rank test and that of a log-rank after removing the censored observations, each at a nominal level of $\alpha = 0.05$. Estimates are based on 2000 Monte Carlo iterations.

Population 1	Size	Unif (0.9, 1.2) Power		
		Unif (0.2, 3.1)	Unif (0.3, 3.1)	Unif (0.4, 3.1)
Population 2	Unif (0.9, 1.2)			
Our imputed log-rank test The log-rank after removing censored observations	0.057	0.584	0.697	0.779
	0.054	0.461	0.564	0.647
				Unif (0.5, 3.1)
				0.858
				0.763

Table 2

Empirical size of a log-rank test where censored observations are counted in the “at risk” sets for both populations, $Y_j, j = 1, 2$. The nominal size used was $\alpha = 0.05$, Estimates are based on 2000 Monte Carlo iterations.

Proportion of Population 1			
Size	0.9	0.8	0.7
	1	0.97	0.55
			0.6
			0.08

Table 3

Empirical size of our imputed log-rank tests using $\alpha = .05$ under simulation scheme (i). All standard errors are less than 0.005.

n	Light censoring				Heavy censoring			
	0.25	0.50	ϕ	0.75	0.25	0.50	ϕ	0.75
50	0.045	0.054		0.058	0.053	0.049		0.055
200	0.047	0.046		0.047	0.048	0.043		0.045

Table 4

Empirical size of our imputed log-rank tests using $\alpha = .0.05$ under simulation scheme (ii). All standard errors are less than 0.005.

ρ	n	Light censoring	Moderately heavy censoring
0.5	50	0.056	0.053
	200	0.054	0.049
0	50	0.054	0.056
	200	0.047	0.051
- 0.5	50	0.047	0.061
	200	0.049	0.051