Published in final edited form as:

Comput Stat Data Anal. 2016 March 1; 95: 150–160. doi:10.1016/j.csda.2015.10.001.

Comparing conditional survival functions with missing population marks in a competing risks model

Dipankar Bandyopadhyay^{a,*} and M. Amalia Jácome^b

^aDepartment of Biostatistics, Virginia Commonwealth University, Richmond, VA, USA

^bMODES Research Group, Faculty of Science, University of A Coruña, Spain

Abstract

In studies involving nonparametric testing of the equality of two or more survival distributions, the survival curves can exhibit a wide variety of behaviors such as proportional hazards, early/late differences, and crossing hazards. As alternatives to the classical logrank test, the weighted Kaplan-Meier (WKM) type statistic and their variations were developed to handle these situations. However, their applicability is limited to cases where the population membership is available for all observations, including the right censored ones. Quite often, failure time data are confronted with missing population marks for the censored observations. To alleviate this, a new WKM-type test is introduced based on imputed population marks for the censored observations leading to fractional at-risk sets that estimate the underlying risk for the process. The asymptotic normality of the proposed test under the null hypothesis is established, and the finite sample properties in terms of empirical size and power are studied through a simulation study. Finally, the new test is applied on a study of subjects undergoing bone marrow transplantation.

Keywords

Competing risk; Fractional risk set; Logrank test; Right censoring; Weighted Kaplan-Meier

1. Introduction

In medical and epidemiological follow-up studies, testing equality of two (or more) survival curves is a commonplace in understanding the effectiveness of treatment or exposure on the corresponding populations. Under the basic assumption of independent samples from two or more populations (including those whose failure times are right-censored), one resorts to the usual rank-based tests, which encompass the logrank (LR) (Mantel, 1966; Cox, 1972; Tarone and Ware, 1977), and the generalized Wilcoxon (Gehan, 1965; Breslow, 1970; Peto and Peto, 1972) statistics, or a Cox's regression (Klein and Moeschberger, 2003). In a

*Address for correspondence: Dipankar Bandyopadhyay, Department of Biostatistics, Virginia Commonwealth University, 830 East Main Street, One Capitol Square, 7th Floor, PO Box 980032, Richmond, VA 23298-0032, USA. Tel: +1 804 827 2058; Fax: +1 804 828 8900; dbandyop@vcu.edu.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

typical competing risk framework where a subject in a healthy state may die due to one of J different risks, one might be interested in comparing the sub-populations (Bandyopadhyay and Datta, 2008) corresponding to the different death types. For example, targeted cancer therapies and vaccine trials devised to reduce the hazard of a specific disease mortality are not expected to affect the mortality from other causes or diseases, so the survival of the individuals dying from that specific disease is expected to be higher than those failing from other causes. However, the setup is complicated by the fact that the marks indicating the cause of death remain unavailable for subjects who were right censored, say, remaining alive at study completion. The cause of death can only be found (via. autopsy or otherwise) after the subject actually dies. In addition to the right censored subjects, determination of the cause of death might be prohibitive for some subjects, thereby contributing to missing marks (Andersen et al., 1996). Removing those incomplete observations would maintain the correct size in the resulting test, however may lead to a loss of power (Bandyopadhyay and Datta, 2008).

Under a competing risk (or multistate) framework, there is a thick body of literature that focuses on estimation of certain (survival) quantities of interest, such as cause-specific hazard (CSH), cumulative incidence function (CIF), as well as regression functions. One of the basic assumptions here is that each subject is clearly identified to belong to one of the independent populations under consideration, including the right-censored ones. Hence, these functions may not be appropriate for summary probabilities of these competing causes. For example, in any clinical trial of a particular cancer type, the patients may die either out of the *cancer* or *other causes*, with a proportion experiencing censoring, say at study completion. The CIF can estimate the chance of a random patient dying due to the cancer before a pre-specified age, say 60. However, it does not answer the question: ‘Among all patients dying due to the cancer (which also includes a proportion of the censored patients eventually experiencing cancer death), what is the probability that a random patient will die before age 60?’. Hence, the correct quantity to estimate and subject to hypothesis testing in this case is the *conditional survival function*, conditional on the cause of failure (Bandyopadhyay and Jácome, 2010). For this, the study subjects/individuals need to be split into J sub-populations, according to their eventual cause of failure. A more formal definition of this problem is available in Section 2.

Under missing population marks, a limited number of papers have dealt with this two-sample problem. Goetghebeur and Ryan (1990) derived a modified logrank test for a two-group comparison problem, while Dewanji (1992) provided a modification to that approach. Using a partial likelihood constructed under semiparametric assumptions, Goetghebeur and Ryan (1995) proposed a score test. Tsiatis et al. (2002) studied a combined logrank test assuming a logistic model for the conditional probability of dying out of the cause of interest and multiple imputation techniques (see e.g. (Rubin, 1987)) to impute missing causes of failure. The aforementioned approaches are based on several parametric assumptions, and therefore there is a clear risk of misspecification of the parametric model. More recently, Bandyopadhyay and Datta (2008) proposed a nonparametric weighted logrank (WLR)-type test for testing conditional survival functions, adapted to the missing population marks setup. Their approach is based on the fractional risks sets (FRS) proposition. The idea of

FRS is to assign fractional probability masses to the censored observations (that has missing population marks) that represent the probability of belonging to each sub-population. That estimate is computed using the nonparametric maximum-likelihood estimator (NLMLE) of the transition probability (further details appear in Section 2). Bandyopadhyay and Datta (2008) showed that the performance of their WLR-FRS test is comparable to the classical WLR test when the population marks are known; however, for missing population marks, the WLR-FRS test outperforms the classical WLR test applied by throwing away the censored observations, while maintaining the size under the null.

The WLR test (both the classical and the FRS versions) is inarguably the most efficient test under the local alternative of proportional hazards (PH) among the two survival functions under comparison. However, it is not always sensitive to the stochastic ordering alternatives, particularly for crossing hazards. In practice, the assumption of a PH between the study and the control groups is unlikely to be true. For example, in surgical interventions, treated patients usually show less favorable short-term results versus placebo but far better long-term results; in the comparison of high/low doses in cancer chemotherapy, high dose may be ineffective initially, but produces favorable long-term results; or in cancer screening trials during a long follow-up period, variations can affect the shape of the observed differences between control and intervention groups (Shapiro et al., 1988). Considering these disadvantages of the WLR test, Pepe and Fleming (1989) introduced the weighted Kaplan-Meier (WKM) statistics as an alternative to the rank-based methods. The WKM test is based on the integrated weighted differences of the Kaplan-Meier (Kaplan and Meier, 1958) estimators of the corresponding survival curves. These WKM statistics are more sensitive to the magnitude of the survival differences than the rank-based tests, compare extremely well with the WLR test, and may perform far better than the WLR test under the crossing hazard alternatives. The asymptotic properties of the WKM test can be found in Pepe and Fleming (1991), and some extensions of the WKM test have been investigated by many authors, including Murray (2001); Shen and Cai (2001); Chi (2005); Lee et al. (2008) and Lee (2011), among others.

In order to alleviate the drawbacks of the WLR tests (both classical and FRS), this paper applies the FRS technique to the WKM test to accommodate missing causes of failure. Thus, the additional contribution is construction of a test where the WLR-FRS test of Bandyopadhyay and Datta (2008) becomes inefficient. The remainder of the paper is organized as follows. In Section 2, the new WKM-FRS test is introduced, and its asymptotic properties are studied. The finite sample performance of the test are explored and compared to the WLR-FRS test in terms of size and power for a variety of alternatives in Section 3. In Section 4, the proposed test is applied to a real dataset on patients undergoing bone marrow transplantation (BMT). Finally, some concluding comments are presented in Section 5 followed by an Appendix containing proofs of the theorems introduced in Section 2.

2. The WKM-FRS test

2.1. Background and Notation

Consider a simple competing risks framework where a set of n subjects are exposed to $J = 2$ competing causes of failure/death, and let $L_j, j = 1, 2$ represent their (latent) failure times,

that is, the potential survival times in hypothetical conditions where the only possible risk is the j th cause. Assuming no censoring, the observable random variables are (T, X) , where $T = \min(L_1, L_2)$ is the actual time to death with survival function $S(t) = P(T > t)$, and the cause of death $X = j$ if $T = L_j$, $j = 1, 2$, i.e., the membership indicator of the two populations to be compared. Let C denote the random censoring time, independent of T and X , with cumulative distribution functions (cdf) $G(t) = P(C \leq t)$ and the conditional cdf $G_j(t) = P(C \leq t | X = j)$, $j = 1, 2$. Under random right censoring, the data actually observed are $\{(Z_i, \delta_i), i = 1, \dots, n\}$, where $Z = \min(T, C)$ is the right-censored failure time with cdf $H(t) = P(Z \leq t)$, and $\delta = \mathbf{1}(T < C)$ is the failure type ($\delta = 1, 2$), or the censoring ($\delta = 0$) indicator. Note that δ_i is observed for each individual, while X_i is observed only for the uncensored ones, with $X_i = \delta_i$ for such cases.

The (conditional) survival function of interest corresponding to the j th cause of failure is

$$S_j(t) = P(T > t | X = j), j = 1, 2 \quad (1)$$

which is the probability of an individual who has survived up to time t eventually dying due to the j th risk, when exposed to the $J = 2$ competing causes of failure. The null hypothesis of interest is that there are no differences in the survival functions S_j between the two competing risks, i.e., $H_0 : S_1 = S_2 (\equiv S(t), \text{ say})$, for all $0 \leq t \leq L$, where $L = \tau$ is a finite but large time point. More details on the choice of L and τ appear in Subsection 2.2 in context of the proposed test statistic.

Remark—One major disadvantage of the latent failure times approach in competing risks is the issue of non-identifiability (Tsiatis, 1975). This means that the cumulative incidence function $CIF_j(t) = P(T \leq t, X = j)$ is identifiable, but does not define the joint distribution of the latent failure times $P(L_1 > t_1, L_2 > t_2)$, unless the latent failure times L_1, L_2 are mutually independent. Following Tsiatis (1975), the identifiability of the quantities $S_j(t)$ also defined

as $S_j(t) = 1 - \frac{CIF_j(t)}{P(X=j)}$ under our competing risk framework is preserved, assuming the independence of L_1 and L_2 . Note that S_j is not $P(L_j > t)$, the probability of an individual surviving up to time t , under the hypothetical condition where cause j is the only risk of death.

The proposed test is based on the integrated weighted difference of appropriate estimators of S_j . The statistical procedures devoted to analyzing the survival functions S_j require computing of the quantities Y_j , defined as

$$Y_j(t) = \sum_{i=1}^n \mathbf{1}(Z_i \geq t, X_i = j), \quad (2)$$

which counts the number of individuals surviving up to time t , and eventually failing due to cause j . This assumes that the membership mark X , the cause of failure, is known for all individuals. Then, the KM estimator of the survival functions S_j is defined as:

$$S_j^{KM}(t) = \prod_{Z_i \leq t} \left(1 - \frac{\mathbf{1}(\delta_i = j)}{Y_j(Z_i)} \right), \quad (3)$$

with Y_j in (2). The stochastic process $Y_j(t)$ is predictable (Bandyopadhyay and Datta, 2008), however the censoring mechanism hinders grouping of individuals depending on their failure type, and hence the population marks X are not available for the censored data. As a consequence, Y_j is not computable under right random censoring. The idea of the FRS formulation is to replace the unknown membership indicator $\mathbf{1}(X_i = j)$ with the estimate $\hat{\phi}_{ij}$ of the probability that the i th individual failing due to the j th cause, given by:

$$\hat{\phi}_{ij} = \begin{cases} 1 & \text{if } \delta_i = j \text{ (failed due to cause } j) \\ 0 & \text{if } \delta_i \neq 0, \delta_i \neq j \text{ (failed due to a cause different from } j) \\ \hat{P}_{0j}(Z_i, \infty) & \text{if } \delta_i = 0 \text{ (censored at time } Z_i), \end{cases}$$

where $\hat{P}_{0j}(s, t)$ is the NPMLE, or the Aalen-Johansen estimator (see Section IV.4.1.1 in Andersen et al., 1993) of the transition probability $P_{0j}(s, t) = P(T - t, X = j | T > s)$, i.e., the probability of an individual who is alive at time s experiencing failure due to cause j by time t , defined as:

$$\hat{P}_{0j}(s, t) = \sum_{s < Z_i \leq t} \left\{ \prod_{s < Z_j < Z_i} \left(1 - \frac{\Delta N(Z_j)}{Y(Z_j)} \right) \right\} \frac{\Delta N_j(Z_i)}{Y(Z_i)},$$

where $N_j(t) = \sum_{i=1}^n \mathbf{1}(Z_i \leq t, \delta_i = j)$ is the number of observed failures of type j in the time interval $[0, t]$, $N(t) = N_1(t) + N_2(t)$ is the number of observed failures irrespective of failure

types in the time interval $[0, t]$ and $Y(t) = \sum_{i=1}^n \mathbf{1}(Z_i \geq t)$ is the size of the at-risk set irrespective of failure types. Then, the FRS estimator of Y_j in (2) is defined as:

$$Y_j^f(t) = \sum_{i=1}^n \mathbf{1}(Z_i \geq t) \hat{\phi}_{ij}. \quad (4)$$

Thus, $Y_j^f(t)$ partitions the risk set $Y(t)$ into J distinct risk sets corresponding to the subpopulations under comparison, and is composed of the number of subjects who died from the j th cause that were alive at time t plus a fractional mass from the censored pool computed using the estimated probability $\hat{P}_{0j}(Z_i, \infty)$. In fact, the fractional masses also determine the estimated proportion of the sample which fails due to cause j . As a consequence, the unknown sample size for the j th subpopulation can be estimated as

$\hat{n}_j = \sum_{i=1}^n \hat{\phi}_{ij}$ (see Satten and Datta, 1999), and the proportion of failures due to cause j can be estimated as $\pi_j = \hat{n}_j / n$. Some theoretical properties of Y_j^f as an appropriate estimator of the at-risk set Y_j have already been studied in the literature. For example, Satten and Datta

(1999) showed that $Y_j^f(t)/Y_j(t) \xrightarrow{P} 1$, and Bandyopadhyay and Jácome (2010) proved that $\sup n^{-1/2}(\ln n)^{-1/2}|Y_j^f(t) - Y_j(t)| \rightarrow 0$.

Under missing population marks, an estimator of the survival function S_j is naturally obtained from the KM estimator in (3) replacing the at-risk set Y_j with the estimated fractional-risk set (4):

$$\hat{S}_j^f(t) = \prod_{Z_i \leq t} \left(1 - \frac{\mathbf{1}(\delta_i=j)}{Y_j^f(Z_i)} \right). \quad (5)$$

The FRS estimators of S_j and the corresponding estimator of the cumulative hazard function Λ_j have been studied by Bandyopadhyay and Jácome (2010), who established some asymptotic properties including uniform strong consistency. The practical performance of these FRS estimators were also explored using both simulations and an application to a real dataset.

2.2. The proposed WKM-FRS test

Based on the integrated weighted difference of the FRS estimators of S_j , the WKM-FRS type statistic is defined as:

$$\Delta(L) = \sqrt{\frac{\hat{n}_1 \hat{n}_2}{n}} \int_0^L W(t) \left(\hat{S}_1^f(t) - \hat{S}_2^f(t) \right) dt,$$

where $\hat{n}_j = \sum_{i=1}^n \hat{\phi}_{ij}$ is the estimated sample size of the j th group, L is a fixed end point, W is data-dependent weight function, and $\hat{S}_j^f$ is the FRS estimator of S_j given in (5). The endpoint L must be chosen such that for $t \leq L$, the differences $\hat{S}_j^f(t) - S_j(t)$, $j=1, 2$ do not exhibit explosive behavior, a feature often observed at the upper tail of the distribution in censored data analysis. The endpoint used by the classical logrank test is an estimator of $\tau_{\min} = \min(\tau_1, \tau_2)$ with $\tau_j = \sup\{t : \min(S_j(t), G_j(t)) > 0\}$ defined as the largest observation time in the j th sample. So $[0, \tau_{\min})$ is the time interval during which subjects in both samples are at risk. The truncation of all data at the smallest of the two largest observations from the two samples yields a loss of information, inappropriate in a non-rank based test statistic. Hence, the choice of L is as follows: estimate the distribution functions S_j and G_j using the FRS estimators $\hat{S}_j^f$ in (5) and

$$\hat{G}_j^f(t) = \prod_{Z_i \leq t} \left(1 - \frac{\mathbf{1}(\delta_i=0)}{Y_j^f(Z_i)} \right), j=1, 2 \quad (6)$$

respectively. Let $\hat{\tau}_j = \sup\{t: \min(\hat{S}_j^f(t), \hat{G}_j^f(t)) > 0\}$, and define $\hat{\tau}_{\min} = \min(\hat{\tau}_1, \hat{\tau}_2)$ and $\hat{\tau}_{\max} = \max(\hat{\tau}_1, \hat{\tau}_2)$. If $\hat{\tau}_{\min}$ is censored, (in which case it may not be possible to compare S_1 and S_2 thereafter), $L = \tau_{\min}$, otherwise $L = \tau_{\max}$ (see Pepe and Fleming, 1991).

The variance of the unweighted test statistic, however, is unstable for L close to the endpoint of the study period. To guarantee stability, the weight function W should downweight the variance of the test at the end of the study period where censoring is heavy. Natural weights involve estimators of the underlying censoring distribution functions. For example, Pepe and Fleming (1989) suggested the weight function:

$$\frac{\hat{G}_1(t-)\hat{G}_2(t-)}{p_1\hat{G}_1(t-)+p_2\hat{G}_2(t-)}, \quad (7)$$

with $\hat{G}_j$ the KM estimator of the censoring distribution functions $G_j, j = 1, 2$, and p_j the observed proportion of data failing due to cause j . When there is no censoring, $\hat{G}_1(t-) = \hat{G}_2(t-) = 1$ for any t , and therefore the weight (7) is 1; so the test can be regarded as a generalization to censored data of the two-sample t -test. On the other hand, the weight (7) downweights the variance of the test at the end of the study period (Pepe and Fleming, 1989, 1991) where censoring is heavy. The WKM test with the weight function in (7) is a competitor to the classical logrank test when the hazards are proportional, and may perform better than the logrank test for crossing hazards alternative. With missing population marks, we propose to use the analogous weight function:

$$W(t) = \frac{\hat{G}_1^f(t-)\hat{G}_2^f(t-)}{\hat{\pi}_1\hat{G}_1^f(t-)+\hat{\pi}_2\hat{G}_2^f(t-)}, \quad (8)$$

where $\hat{G}_j^f$ in (6) is the FRS estimator of G_j and $\hat{\pi}_j = n^{-1} \sum_{i=1}^n \hat{\phi}_{ij}$ is the estimated proportion of failures due to cause j .

The asymptotic distribution of (L) is derived under the null hypothesis $H_0 : S_1 = S_2$. Because the FRS are used to handle the unknown sub-population marks, we cannot use the standard counting process formulation and the related martingale techniques since the imputed marks $\hat{\phi}_{ij}$ are based on the entire data set, and the fractional at-risk sets Y_j^f are not predictable with respect to the natural filtration process. Nevertheless, we present the asymptotic normality of (L) under the null by writing the test statistic as a sum of independent and identically distributed (iid) mean zero variable plus a remainder term converging to zero in probability.

Further notations required to describe the iid representation of the proposed test involves the following functions:

$$\overline{H}^c(t) = P(Z \geq t, \delta=0) \text{ and } H^{nc}(t) = P(Z \leq t, \delta>0),$$

for the censored ($\delta = 0$) and not censored ($\delta > 0$) data, respectively, as well as the following functions of the data failing due to cause j :

$$\overline{H}_j^{nc}(t) = P(Z \leq t, \delta = j) \text{ and } \overline{H}_j(t) = P(Z \geq t, X = j).$$

Let $I[f(t)] = \int_t^L f(u) du$ for a function f . We will assume that W and ω satisfy the following conditions:

Assumption 1—The predictable weighting process $W(t)$ is a suitable estimator of the nonstochastic function $w(t)$ such that

$$\sup_{t \in [0, L)} |W(t) - \omega(t)| \xrightarrow{P} 0.$$

Assumption 2—The function ω is piecewise continuous with finitely many discontinuous points. Except on this finite set of discontinuity points, we have $I[W(t)] \xrightarrow{P} I[\omega(t)]$ for all $t \in [0, L]$ as $n \rightarrow \infty$.

Assumption 3—Let $V_{[0, L]}(f)$ denote the local variation of f in $[0, L]$. Then, $V_{[0, L]}(I[W(t)]) + V_{[0, L]}(I[\omega(t)])$ is bounded in probability.

Under these assumptions, we have the following theorems on the linear representation of the WKM-FRS statistic, and its asymptotic normality under H_0 .

Theorem 2.1—Let π_j be the proportion of individuals in the j th group and ω the in-probability limit of the weight function W . Under Assumptions 1–3, if S_j and G_j , $j = 1, 2$ are continuous, then under the null hypothesis $H_0: S_1 = S_2 \equiv S$, the WKM-FRS test can be written as the following linear representation:

$$\Delta(L) = n^{-1/2} \sum_{i=1}^n \Delta_i^*(L) + o_P(1), \quad (9)$$

with

$$\Delta_i^*(L) = \sqrt{\pi_1 \pi_2} \int_0^L \omega(t) S(t) [\xi_2(Z_i, X_i, \delta_i, t) - \xi_1(Z_i, X_i, \delta_i, t)] dt, \quad (10)$$

where

$$\xi_j(Z, X, \delta, t) = \frac{\mathbf{1}(Z \leq t, X = j)}{\overline{H}_j(Z)} - \int_0^t \frac{\mathbf{1}(Z \leq v, X = j)}{\overline{H}_j^2(v)} dH_j^{nc}(v) + \int_0^t \frac{\rho_j(Z, X, \delta, v)}{\overline{H}_j^2(v)} dH_j^{nc}(v),$$

$$\rho_j(Z, X, \delta, v) = \mathbf{1}(Z \geq v, \delta=0) [\mathbf{1}(X=j) - P_j(Z, \infty)] - \int_v^\infty \zeta_j(Z, X, \delta, u) \frac{d\bar{H}^c(u)}{S(u)},$$

and

$$\begin{aligned} \zeta_j(Z, X, \delta, u) = & \frac{S(Z)}{1 - H(Z)} \mathbf{1}(Z \geq u, \delta > 0) [\mathbf{1}(X=j) - P_j(Z, \infty)] \\ & - \int_u^\infty \frac{S(z)}{(1 - H(z))^2} \mathbf{1}(Z \geq z) [\mathbf{1}(X=j) - P_j(z, \infty)] dH^{nc}(z). \end{aligned}$$

An outline of the proof of the representation (9) is given in the Appendix.

Theorem 2.2—Under Assumptions 1–3, if S_j and G_j , $j = 1, 2$ are continuous, then the FRS-WKM test $\Delta(L)$ converges in distribution under the null hypothesis $H_0 : S_1 = S_2 \equiv S$ to a zero-mean normal density given by:

$$\Delta(L) \xrightarrow{D} N(0, \sigma^2(L)),$$

with

$$\begin{aligned} \sigma^2(L) = & \int_0^L \int_0^L \omega(t_1) \omega(t_2) S(t_1) S(t_2) \int_0^{t_1 \wedge t_2} \frac{dH^{nc}(v)}{(1 - H(v))^2} dt_1 dt_2 \\ & + \int_0^L \int_0^L \omega(t_1) \omega(t_2) S(t_1) S(t_2) \int_0^{t_1} \int_0^{t_2} g(v_1, v_2) \frac{dH^{nc}(v_1) dH^{nc}(v_2)}{(1 - H(v_1))^2 (1 - H(v_2))^2} dt_1 dt_2, \end{aligned} \quad (11)$$

where

$$g(v_1, v_2) = \int_{v_1}^\infty \int_{v_2}^\infty \int_{x \vee y}^\infty \left(\frac{S(z)}{1 - H(z)} \right)^2 dH^{nc}(z) \frac{dH^c(x) dH^c(y)}{S(x) S(y)}$$

and ω is the in-probability limit of the weight function W .

Let $\hat{\sigma}^2(L)$ be a pooled consistent estimator of $\sigma^2(L)$ under the null hypothesis. Then, under the conditions of Theorem 2.2, we have $\Delta(L)/\hat{\sigma}(L) \xrightarrow{D} N(0, 1)$. Therefore, a two-sided α -level test rejects H_0 whenever $|\Delta(L)/\hat{\sigma}(L)| > z_{\alpha/2}$, where z_α is the $(1-\alpha) \times 100$ -th percentile of the standard normal distribution. Under the null hypothesis, the estimator $\hat{\sigma}^2(L)$ can be easily obtained using pooled estimates of the functions in the asymptotic closed-form variance (11) as follows:

$$\begin{aligned}
\hat{\sigma}^2(L) = & 2 \sum_{Z_k \leq L} \sum_{Z_j \leq Z_k} \sum_{Z_i \leq Z_j} W(Z_k) W(Z_j) \frac{S_n^{KM}(Z_k) S_n^{KM}(Z_j)}{(1 - H_n^2(Z_i))} \mathbf{1}(\delta_i > 0) (Z_j - Z_{j-1}) (Z_k - Z_{k-1}) \\
& + \frac{1}{n^2} \sum_{\substack{Z_k \leq L \\ Z_i \leq L, Z_j \leq Z_k}} W(Z_i) W(Z_k) \frac{S_n^{KM}(Z_i) S_n^{KM}(Z_k)}{(1 - H_n^2(Z_i))(1 - H_n^2(Z_j))} \hat{g}(Z_i, Z_j) \\
& \times \mathbf{1}(\delta_i > 0, \delta_j > 0) (Z_k - Z_{k-1}) (Z_i - Z_{i-1}),
\end{aligned} \tag{12}$$

with

$$\hat{g}(v_1, v_2) = \frac{1}{n^3} \sum_{\substack{Z_i \geq v_1 \\ Z_j \geq v_2}} \sum_{Z_k \geq Z_i \vee Z_j} \frac{1}{S_n^{KM}(Z_i) S_n^{KM}(Z_j)} \left(\frac{S_n^{KM}(Z_k)}{1 - H_n(Z_k)} \right)^2 \mathbf{1}(\delta_i > 0, \delta_j > 0, \delta_k > 0),$$

where W is the weight function (8), S_n^{KM} the pooled KM estimator of S , and H_n the empirical estimator of H . The asymptotic consistency of $\hat{\sigma}^2(L)$ as a suitable estimator of $\sigma^2(L)$ can be established through the consistency properties of the aforementioned estimators (the theoretical details are not presented here).

The empirical estimator of $\sigma^2(L)$ given in (12) is rather involved and unwieldy from a computational point of view. As an alternative, we propose a bootstrap method to approximate the variance of the WKM-FRS test statistic (L) under the null hypothesis. The bootstrap resampling plan follows that in Bandyopadhyay and Datta (2008):

- i. Compute $d_i = \mathbf{1}(\delta_i > 0)$, $i = 1, \dots, n$, the true failure indicators irrespective of the failure types. Then, generate iid resamples $\{(Z_i^*, d_i^*), i=1, \dots, n\}$ from the empirical distribution of the pairs $\{(Z_i, d_i), i = 1, \dots, n\}$.
- ii. Generate the indicators δ_i^* , $i = 1, \dots, n$ as follows:
 - If $d_i^* = 0$ then $\delta_i^* = 0$
 - If $d_i^* = 1$ then let $\delta_i^* = j$ with probability $P(\delta_i^* = j) = \hat{n}_j / n$ where $\hat{n}_j = \sum_{i=1}^n \hat{\phi}_{ij}$.
- iii. With the bootstrap sample $\{(Z_i^*, \delta_i^*), i=1, \dots, n\}$, compute the bootstrap test statistic $^*(L)$.
- iv. Repeat steps (i)–(iii) a large number of times to obtain B values of the corresponding test statistics $\{\Delta_b^*(L), b=1, \dots, B\}$
- v. A bootstrap estimate of the variance of (L) is given by the empirical variance of $\{\Delta_b^*(L), b=1, \dots, B\}$.

Note that this bootstrap resampling scheme adapts the simple bootstrap method for censored data described in Efron (1981) to the missing population marks setting. This was also successfully implemented in Bandyopadhyay and Datta (2008). The asymptotic validity (consistency) of this bootstrap based estimator of the variance of (L) can be established by

deriving a similar linear representation as (9) for the resampled statistic. This is currently omitted for brevity.

3. Simulation study

In this section, a simulation study was conducted to assess the size and power properties of the WKM-FRS statistic for finite samples, assuming a competing risk framework with $J = 2$ cause of failure. Data were generated under five different scenarios, which included the null and four alternative cases (Cases I–IV, described later). The results of the proposed test were evaluated for small ($n_1 = n_2 = 50$) and moderate ($n_1 = n_2 = 100$) sample sizes. The power was computed via Monte Carlo simulations, and defined as the proportion of rejections out of $m = 1000$ replications for each configuration. The variance of the proposed test was estimated using the bootstrap method described in Section 2, with a total of $B = 1000$ bootstrap resamples. For comparison, we also considered the WLR-FRS test by Bandyopadhyay and Datta (2008). All tests were two-sided considering a nominal type I error level $\alpha = 0.05$.

3.1. Size

The failure times L_1 and L_2 corresponding to the two competing causes of failure were generated independently as two exponential distributions $\text{Exp}(\lambda_1)$ and $\text{Exp}(\lambda_2)$ respectively, with $\lambda_1 = \lambda_2 = 1$. The censoring distribution was selected independently of the failure times as uniform $U(0, 5)$ and $U(0, 2)$, representing 20% (light) and 40% (heavy) censoring rates, respectively. The empirical sizes obtained from the asymptotic distribution of both FRS-based tests were fairly close to the nominal 5% level (see Table 1).

3.2. Power

For the alternative configurations, four distinct patterns were selected, namely proportional hazards (Case I), early differences (Case II), late differences (Case III), and crossing hazards (Case IV). While the failure times were generated from piecewise exponential distributions, the censoring times follow $U(0, c)$ distributions, with c chosen to represent approximately 20% (light) and 40% (heavy) censoring. The parameter values considered were: Case I, $\lambda_1 = 2, \lambda_2 = 1, c = 1.5$ (heavy) and $c = 4$ (light); Case II, $\lambda_1 = 1, 0.75, \lambda_2 = 0.25, 0.75$ for $t < 1, 1 \leq t$, respectively, $c = 8$ (light), $c = 3.5$ (heavy); Case III, $\lambda_1 = 0.25, 0.25, \lambda_2 = 0.33, 0.05$ for $t < 4, 4 \leq t$, respectively, $c = 24$ (light), $c = 8.5$ (heavy); Case IV, $\lambda_1 = 2.0, 0.5, 0.5, \lambda_2 = 0.5, 0.5, 2.0$ for $t < 0.5, 0.5 \leq t < 0.6, 0.6 \leq t$, respectively, $c = 4.5$ (light), $c = 1.9$ (heavy). Similar configurations were previously considered in Pepe and Fleming (1989), or Lee (2011) among others.

The simulation results are presented in Table 2. They agree with the comparisons between the LR and KM type tests in the literature with complete population marks (see Pepe and Fleming, 1989; Lee et al., 2008, among others). The WLR tests are known to be the most powerful tests in detecting PH alternatives. So, for Case I, the WLR-FRS test performed better than the WKM-FRS across all sample sizes and censoring levels, though the difference was small. This indicates that the proposed WKM-FRS test attains reasonable efficiency against proportional hazards. For Case II with early survival differences, the

proposed test was superior to the WLR-FRS test in small and moderate sample sizes at all censoring percentages. The difference in power was considerable for light censoring (48% higher for $n = 50$), as compared to the heavy (33% higher $n = 50$). For Case III with late survival differences, the WLR-FRS test appeared to be preferable over the proposed test, although the differences in power was not substantial especially with light censoring (14% and 20% for $n = 50$ and $n = 100$, respectively). Finally, for Case IV with crossing hazards, the proposed test outperformed the WLR-FRS test substantially, particularly for low censoring (138% and 166% higher power for $n = 50$ and $n = 100$ respectively). Interestingly, for crossing hazards alternatives, the power of the WLR-FRS test increased for higher censoring, a feature also observed in Lee et al. (2008). In particular, power is a complicated functional that not only depends on the alternative, but also on the censoring distribution and the point of crossing of the two survival curves. Usually, under right censoring, early survival differences are mainly observed under heavy censoring. However, this feature of increased power may also very well reflect the fact that the choice of c for the censoring density $\text{Uniform}(0, c)$ induces a wider separation between the two survival curves under increased censoring, leading to a larger functional and consequently higher power.

4. Application

For illustration, the proposed WKM-FRS test and the WLR-FRS test were applied to a real data example consisting of 300 bone marrow transplant patients with age 50 years, or more from the BMT program at University of Minnesota Medical Center. The transplantation dates were between 2001–2013. These patients underwent transplantation for various disease types, which includes acute myeloid leukemia (AML), acute lymphocytic leukemia (ALL), and myelodysplasia (MDS), and includes both non-myeloablative and myeloablative conditioning from allogenic donors (donors that are either related or unrelated to the patients). Among these patients who received transplants, about 35% experienced relapse (R) typically leading to death (Cause 1), 28% had death unrelated to relapse (UR, Cause 2), and the rest 37% were right-censored observations whose eventual cause of death remained unknown till study completion. Time to event (relapse/nonrelapse death), or the last follow-up was recorded. The failure times ranged from 7 days to 3536 days, with the R patients from 29 to 3536 days and the UR patients from 7 to 2808 days.

The plots of the estimated FRS-based survival functions $\hat{S}_j^f$ for the two competing causes are presented in Figure 1. Using $B = 5000$ bootstrap replicates to compute the standard error of the test statistic, the p -value for the two-sided alternative was 0.0002 for the WKM-FRS test and 0.062 for the WLR-FRS test. Therefore, the null hypothesis of no survival difference in the two groups is not rejected at the 5% level by the WLR-FRS test, while we can conclude that the survival functions corresponding to the two failure types are significantly different according to the WKMFRS test. From Figure 1, it is clear that the two survival functions cross each other, so that the early survival differences in favor of the relapse group are negated by the late survival differences for the non-relapse death group, leading to a poor behavior of the WLR-FRS test. Furthermore, testing with the WKM-FRS test, for a one-sided alternative reveals that the non-relapse patients (Cause 2) have a shorter survival time than those who died due to relapse (Cause 1).

5. Conclusions

In this paper, a new class of test statistic was constructed for comparing (conditional) survival functions as an alternative to the usual rank-based tests. To handle the missing causes of failure for randomly right-censored subjects, the fractional risks set (FRS) technique was considered. Some asymptotic properties of the test statistic were studied and a bootstrap scheme was proposed for computing the standard errors. Simulation studies and application to a real dataset revealed that the proposed test outperforms the analogous FRS-based weighted log-rank statistics of Bandyopadhyay and Datta (2008), in particular for crossing hazards.

The proposed WKM-FRS test can be used to test for the dependence between the failure time T and the cause of failure X . Under independence, T and X can be studied separately, simplifying inference to a great extent. For example, testing for the equality of competing risks based on their cumulative incidence functions or their cause specific hazard rates reduces to testing the null hypothesis $H_0 : \pi_1 = \pi_2$. This implies that, under H_0 , each uncensored time can be randomly attributed to one of the two possible causes of failure. There is only a limited number of papers dealing with the testing problem of independence between T and X , and all of them ignore censoring. While some tests addressed the cause specific hazard rates (Aly et al., 1994; Dykstra et al., 1998), the others considered conditional probabilities of the form $P(X = j|T = t)$ (see Kochar and Proschan, 1991), or $P(X = j|T > t)$ (see Dewan et al., 2004; Gasbarra et al., 2006). Under right censoring, the difficulty relays in how to handle the subjects with missing values for X . Once again, based on the FRS technique, the independence between T and X can be tested by considering the null hypothesis of no difference between the two conditional survival functions $S_j(t) = P(T > t|X = j)$, $j = 1, 2$.

The proposed WKM-FRS test can be improved to attain higher power against certain alternatives by carefully choosing the weight function W . For example, one might be tempted to incorporate the Fleming and Harrington class of weight functions $G^{\rho, \gamma}$, into W (see Fleming and Harrington, 1991), leading to a new weight function given by $W^{\rho, \gamma}(t) = G^{\rho, \gamma}(t)W(t)$ with W defined in (8), $G^{\rho, \gamma}(t) = \hat{S}(t-)^{\rho} \{1 - \hat{S}(t-)\}^{\gamma}$ for $\rho \geq 0$, $\gamma \geq 0$, and $\hat{S}(\cdot)$ the Kaplan-Meier estimator in the pooled sample. Optimal combinations of the values of ρ and γ would ascertain maximum power for arbitrary alternatives. Starting with this new weight function, one can develop a class of versatile tests by considering stochastic version of the WKM statistics (see Lee et al., 2008), maximum of the WKM tests statistics over the different weights (see Shen and Cai, 2001; Lee, 2011), or comparisons of paired censored survival lifetimes (Murray, 2001). These are avenues for future research, and will be considered elsewhere.

Acknowledgements

The authors thanks the Associate Editor and two anonymous reviewers whose insightful comments led to a considerable improved version of the manuscript. They also thank Prof. Somnath Datta for suggesting the problem, and Todd DeFor from the University of Minnesota Masonic Cancer Center for providing the BMT dataset. Bandyopadhyay acknowledges support from the US National Institutes of Health grants R03DE023372 and R01DE024984. M. Amalia Jácome's research was supported in part by grants MTM2011-22392 and CN 2012/130.

6. Appendix

Proof. of Theorem 2.1

The decomposition of the proposed WKM-FRS test as a sum of iid mean zero variables plus a negligible term is based on the iid representation of the FRS-based estimator of the cumulative hazard function in Bandyopadhyay and Jácome (2010). The key feature is that the estimated proportions of failure for the j th cause π_j can be replaced by the populational proportions $\pi_j, j = 1, 2$ and the random weight function W by its nonrandom limit ω , without affecting the asymptotic distribution of $\Delta(L)$. Then, the WKM-FRS test can be split into three terms where one of them equals zero under the null and another negligible in probability. Finally, the dominant term admits an iid representation, following Theorem 3 of Bandyopadhyay and Jácome (2010).

From Theorem 3.1 in Datta and Satten (2000) and the weak law of large numbers, we immediately have $\hat{\pi}_j = \hat{\pi}_j/n$ as a weakly consistent estimator of π_j , the proportion of individuals in the j th sub-population, and therefore,

$$\Delta(L) = \tilde{\Delta}(L) + o_p(1) \quad (13)$$

with

$$\tilde{\Delta}(L) = \sqrt{n} \sqrt{\pi_1 \pi_2} \int_0^L W(t) (\hat{S}_1^f(t) - \hat{S}_2^f(t)) dt.$$

Let us define the process

$$B_{nj}(t) = \sqrt{n}(\hat{S}_j^f(t) - S_j(t)), j=1, 2. \quad (14)$$

Now, $\tilde{\Delta}(L)$ can be split into three terms as follows:

$$\tilde{\Delta}(L) = A + B + R_n(L), \quad (15)$$

where $A = \sqrt{\pi_1 \pi_2} \int_0^L \omega(t)(B_{n1}(t) - B_{n2}(t))dt$ contributes to the iid representation of the WKM-FRS test, $B = \sqrt{n} \sqrt{\pi_1 \pi_2} \int_0^L W(t)(S_1(t) - S_2(t))dt$ equals zero under the null hypothesis $H_0 : S_1 = S_2$, and the remaining term is

$$R_n(t) = \sqrt{\pi_1 \pi_2} \int_0^L (W(t) - \omega(t))(B_{n1}(t) - B_{n2}(t))dt. \quad (16)$$

To prove that R_n is negligible in probability, consider $B_{nj}(t)$ in (14) in terms of the process $L_{nj}(t) = \sqrt{n}(\hat{\Lambda}_j^f(t) - \Lambda_j(t))$, where Λ_j is the cumulative hazard function associated to the conditional survival function S_j , and $\hat{\Lambda}_j^f$ the FRS estimator of Λ_j given by

$$\hat{\Lambda}_j^f(t) = \sum_{Z_i \leq t} \frac{\Delta N_j(Z_i)}{Y_j^f(Z_i)}. \text{ Similarly, as in Breslow and Crowley (1974),}$$

$$\sqrt{n} \left(\hat{S}_j^f(t) - \exp(-\hat{\Lambda}_j^f(t)) \right) \xrightarrow{P} 0. \quad (17)$$

Furthermore, by the delta method, we have

$$\exp(-\hat{\Lambda}_j^f(t)) - \exp(-\Lambda_j(t)) = -S_j(t)(\hat{\Lambda}_j^f(t) - \Lambda_j(t)) + o_p(1). \quad (18)$$

Then, from (17), (18), the iid representation of $L_{nj}(t)$ in Theorem 3 of Bandyopadhyay and Jácome (2010), the SLLN and the delta method, $B_{nj}(t)$ converges weakly to a zero mean Gaussian process. Therefore, under Assumptions 1–3 and applying Lemma 1 of Gu, Follmann, and Geller (1999) to $\int_t^L W(u)du$, the term R_n in (16) is negligible in probability. As a consequence, from (13), (15), (17) and (18), the WKM-FRS test (under the null) takes the form:

$$\Delta(L) = \sqrt{\pi_1 \pi_2} \int_0^L \omega(t) (L_{n2}(t) - L_{n1}(t)) dt + o_p(1).$$

Finally, from the iid representation of $L_{nj}(t)$, $j = 1, 2$ as in Theorem 3 from Bandyopadhyay and Jácome (2010), we have (under the null)

$$\Delta(L) = \sqrt{\frac{\pi_1 \pi_2}{n}} \int_0^L \omega(t) S(t) \sum_{i=1}^n (\xi_2(Z_i, X_i, \delta_i, t) - \xi_1(Z_i, X_i, \delta_i, t)) dt + o_p(1),$$

where ξ_j , $j = 1, 2$ is defined in Theorem 2.1. This completes the proof.

Proof. of Theorem 2.2

As noted in Section 2.2, standard martingale theory cannot be directly applied to prove the asymptotic normality since the proposed test $\Delta(L)$ is not predictable with respect to the natural filtration process. Nevertheless, the asymptotic distribution of $\Delta(L)$ coincides with that of the iid representation, a sum of independent identically distributed mean zero variable :

$$\Delta(L) = n^{-1/2} \sum_{i=1}^n \Delta_i^*(L) + o_p(1)$$

with Δ_i^* given in (10). Under H_0 , the variance of the asymptotic distribution of $\Delta(L)$ is

$$\text{var}(\Delta_1^*(L)) = \pi_1 \pi_2 \int_0^L \int_0^L \omega(t_1) \omega(t_2) S(t_1) S(t_2) (V_{11}(t_1, t_2) + V_{22}(t_1, t_2) - 2V_{12}(t_1, t_2)) dt_1 dt_2,$$

where

$$V_{ij}(t_1, t_2) = E[\xi_i(Z, X, \delta, t_1)\xi_i(Z, X, \delta, t_2)],$$

with ξ_j defined in Theorem 2.1. Under H_0 , some tedious but straightforward calculations yields,

$$V_{jj}(t_1, t_2) = \frac{1}{\pi_j} \int_0^{t_1 \wedge t_2} \frac{dH^{nc}(v)}{(1 - H(v))^2} + \frac{1 - \pi_j}{\pi_j} \int_0^{t_1} \int_0^{t_2} \frac{g(v_1, v_2)}{(1 - H(v_1))^2 (1 - H(v_2))^2} dH^{nc}(v_1) dH^{nc}(v_2),$$

with

$$g(v_1, v_2) = \int_{v_1}^{\infty} \int_{v_2}^{\infty} \frac{m(u_1 \vee u_2)}{(1 - F(u_1))(1 - F(u_2))} dH^c(u_1) dH^c(u_2)$$

and

$$m(u) = \int_u^{\infty} \left(\frac{S(z)}{1 - H(z)} \right)^2 dH^{nc}(z).$$

For V_{12} , standard algebra yields

$$\begin{aligned} V_{12}(t_1, t_2) &= \int_0^{t_1} \int_0^{t_2} E[\rho_1(Z, X, \delta, u)\rho_2(Z, X, \delta, v)] \frac{dH_1^{nc}(u)dH_2^{nc}(v)}{(1 - H_1(u))^2(1 - H_2(v))^2} \\ &= -\pi_1\pi_2 \int_0^{t_1} \int_0^{t_2} g(u, v) \frac{dH_1^{nc}(u)dH_2^{nc}(v)}{(1 - H_1(u))^2(1 - H_2(v))^2}, \end{aligned}$$

with ρ_j and g are defined in Theorems 2.1 and 2.2, respectively. Combining the results, we obtain the expression of the variance $\sigma^2(L)$ given in (11).

References

- Aly EEA, Kochar SC, McKeague IW. Some tests for comparing cumulative incidence functions and cause-specific hazard rates. *Journal of the American Statistical Association*. 1994; 89(427):994–999.
- Andersen J, Goetghebuer E, Ryan L. Missing cause of death information in the analysis of survival data. *Statistics in Medicine*. 1996; 15:2191–2201. [PubMed: 8910963]
- Andersen, PK.; Borgan, O.; Gill, RD.; Keiding, N. *Statistical Models Based on Counting Processes*. Springer; 1993.
- Bandyopadhyay D, Datta S. Testing equality of survival distributions when the population marks are missing. *Journal of Statistical Planning and Inference*. 2008; 138(6):1722–1732. [PubMed: 19844606]
- Bandyopadhyay D, Jácome A. Nonparametric estimation of conditional cumulative hazards for missing population marks. *Australian and New Zealand Journal of Statistics*. 2010; 52(1):75–91. [PubMed: 20717497]
- Breslow N. A generalized Kruskal-Wallis test for comparing k samples subject to unequal patterns of censorship. *Biometrika*. 1970; 57(3):579–594.

- Breslow N, Crowley J. A large sample study of the life table and product limit estimates under random censorship. *The Annals of Statistics*. 1974; 2(3):437–453.
- Chi Y. Multiple testing procedures based on weighted Kaplan-Meier statistics for right-censored survival data. *Statistics in Medicine*. 2005; 24(1):23–35. [PubMed: 15515153]
- Cox DR. Regression models and life-tables. *Journal of the Royal Statistical Society. Series B*. 1972; 34(2):187–220.
- Datta S, Satten GA. Estimating future stage entry and occupation probabilities in a multistage model based on randomly right-censored data. *Statistics and Probability Letters*. 2000; 50(1):89–95.
- Dewan I, Deshpande J, Kulathinal S. On testing dependence between time to failure and cause of failure via conditional probabilities. *Scandinavian Journal of Statistics*. 2004; 31(1):79–91.
- Dewanji A. A note on a test for competing risks with missing failure type. *Biometrika*. 1992; 79(4): 855–857.
- Dykstra R, Kochar S, Robertson T. Restricted tests for testing independence of time to failure and cause of failure in a competing-risks model. *Canadian Journal of Statistics*. 1998; 26(1):57–68.
- Efron B. Censored data and the Bootstrap. *Journal of the American Statistical Association*. 1981; 76:312–319.
- Fleming, TR.; Harrington, DP. *Counting Processes and Survival Analysis*. New York: John Wiley & Sons; 1991.
- Gasbarra D, Kulathinal S, Dewan I, Nissinen A. Testing dependence between the failure time and failure modes: An application of enlarged filtration. *Journal of Statistical Planning and Inference*. 2006; 136(5):1669–1686.
- Gehan EA. A generalized wilcoxon test for comparing arbitrarily singly-censored samples. *Biometrika*. 1965; 52(1–2):203–223. [PubMed: 14341275]
- Goetghebuer E, Ryan L. A modified log rank test for competing risks with missing failure type. *Biometrika*. 1990; 77:207–211.
- Goetghebuer E, Ryan L. Analysis of competing risks survival data when some failure types are missing. *Biometrika*. 1995; 82:821–833.
- Gu M, Follmann D, Geller NL. Monitoring a general class of two-sample survival statistics with applications. *Biometrika*. 1999; 86(1):45–57.
- Kaplan EL, Meier P. Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association*. 1958; 53(282):457–481.
- Klein, J.; Moeschberger, M. *Survival analysis: Statistical methods for censored and truncated data*. New York, NY: Springer-Verlag; 2003.
- Kochar SC, Proschan F. Independence of time and cause of failure in the multiple dependent competing risks model. *Statistica Sinica*. 1991; 1:295–299.
- Lee SH. Maximum of the weighted Kaplan-Meier tests for the two-sample censored data. *Journal of Statistical Computation and Simulation*. 2011; 81(8):1017–1026.
- Lee SH, Lee EJ, Omolo BO. Using integrated weighted survival difference for the two-sample censored data problem. *Computational Statistics and Data Analysis*. 2008; 52:4410–4416.
- Mantel N. Evaluation of survival data and two new rank order statistics arising in its consideration. *Cancer Chemother Reports*. 1966; 50:163–170.
- Murray S. Using Weighted Kaplan-Meier Statistics in Nonparametric Comparisons of Paired Censored Survival Outcomes. *Biometrics*. 2001; 57(2):361–368. [PubMed: 11414557]
- Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: A class of distance tests for censored survival data. *Biometrics*. 1989; 45(2):497–507. [PubMed: 2765634]
- Pepe MS, Fleming TR. Weighted Kaplan-Meier statistics: Large sample and optimality considerations. *Journal of the Royal Statistical Society Series B*. 1991; 53(2):341–352.
- Peto R, Peto J. Asymptotically efficient rank invariant test procedures. *Journal of the Royal Statistical Society Series A*. 1972; 137(2):185–207.
- Rubin, DB. *Multiple Imputation for Nonresponse in Surveys*. New York: Wiley; 1987.
- Satten GA, Datta S. Kaplan–Meier representation of competing risk estimates. *Statistics and Probability Letters*. 1999; 42(3):299–304.

- Shapiro, S.; Venet, W.; Strax, P.; Venet, L. The Health Insurance Plan Project and Its Sequelae, 1963–1986. Baltimore: The Johns Hopkins University Press; 1988. Periodic Screening for Breast Cancer.
- Shen Y, Cai J. Maximum of the Weighted Kaplan-Meier Tests with Application to Cancer Prevention and Screening Trials. *Biometrics*. 2001; 57(3):837–843. [PubMed: 11550935]
- Tarone RE, Ware J. On distribution-free tests for equality of survival distributions. *Biometrika*. 1977; 64(1):156–160.
- Tsiatis A. A nonidentifiability aspect of the problem of competing risks. *Proceedings of The National Academy of Sciences, USA*. 1975; 72(1):20–22.
- Tsiatis AA, Davinian M, McNeney B. Multiple imputation methods for testing treatment differences in survival distributions with missing cause of failure. *Biometrika*. 2002; 89(1):238–244.

Figure 1.

Estimated survival functions using fractional risk sets for the BMT dataset. The two competing causes are (i) Relapse death and (ii) Non-relapse death.

Table 1Size simulation results at $\alpha = 0.05$

$n_1 = n_2$	Censoring	WKM-FRS	WLR-FRS
50	20%	0.048	0.051
	40%	0.042	0.047
100	20%	0.056	0.052
	40%	0.053	0.051

Table 2

Power simulation results at $\alpha = 0.05$ for the cases I–IV.

	$n_1 = n_2$	Censoring	WKM-FRS	WLR-FRS
I. Prop. Hazards	50	20%	0.306	0.391
		40%	0.109	0.136
	100	20%	0.504	0.603
		40%	0.159	0.213
II. Early differences	50	20%	0.739	0.498
		40%	0.543	0.459
	100	20%	0.944	0.708
		40%	0.789	0.596
III. Late differences	50	20%	0.174	0.199
		40%	0.105	0.161
	100	20%	0.218	0.262
		40%	0.143	0.251
IV. Crossing hazards	50	20%	0.739	0.310
		40%	0.674	0.601
	100	20%	0.937	0.352
		40%	0.915	0.753