Published in final edited form as:

Scand Stat Theory Appl. 2017 June ; 44(2): 545–562. doi:10.1111/sjos.12265.

Nonparametric Two-Sample Tests of the Marginal Mark Distribution with Censored Marks

BRENT A. JOHNSON

Department of Biostatistics and Computational Biology, University of Rochester

Abstract

Occasionally, investigators collect auxiliary marks at the time of failure in a clinical study. Because the failure event may be censored at the end of the follow-up period, these marked endpoints are subject to induced censoring. We propose two new families of two-sample tests for the null hypothesis of no difference in mark-scale distribution that allows for arbitrary associations between mark and time. One family of proposed tests is a nonparametric extension of an existing semi-parametric linear test of the same null hypothesis while a second family of tests is based on novel marked rank processes. Simulation studies indicate that the proposed tests have the desired size and possess adequate statistical power to reject the null hypothesis under a simple change of location in the marginal mark distribution. When the marginal mark distribution has heavy tails, the proposed rank-based tests can be nearly twice as powerful as linear tests.

Keywords

copula; induced censoring; log-rank test; marked point process; quantile process; survival analysis; Wilcoxon statistic

1. Introduction

In some studies, investigators have an interest in auxiliary outcomes measured at the occurrence of a failure event. In clinical studies with limited follow-up, the failure event may not be observed due to the end of the study or loss to follow-up. In such cases, censoring on the time-scale induces censoring of the auxiliary, or marked outcome and ordinary methods of survival analysis generally do not apply. Our goal is to propose a new two-sample test that may be applied to a variety of different applications under mild assumptions about the data.

Because of the vast and diverse literature on marked point processes and the statistical analysis of marked endpoints, we clarify some fundamental assumptions about our definition of marked endpoint and, in particular, what is observed and unobserved in our setup. We very briefly review two types of marked outcomes: first, there are marks that present themselves at the random event time T and effectively do not exist prior to time T ; secondly,

Brent A. Johnson, Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Avenue, Box 630, Rochester, NY 14642, USA. brent_johnson@urmc.rochester.edu.

Supporting information

Additional information for this article is available online:

accumulation processes measured at time T . An example of the first type of marked endpoint comes from HIV vaccine studies as described in a series of papers by Gilbert and colleagues (e.g. Gilbert *et al.*, 2004; Gilbert *et al.*, 2008). In HIV vaccine studies, study participants are randomized to two or more treatment groups (e.g. vaccine or no vaccine) and then followed until they become infected with HIV at time T or until they are censored at the end of follow-up. The mark of interest Z is a continuous measure of genetic distance that summarizes the HIV strain and used to assess vaccine efficacy. The second type of marked endpoint is derived from accumulation processes, and two examples include quality-adjusted survival (e.g. Gelber *et al.*, 1989; Glasziou *et al.*, 1990; Zhao & Tsiatis, 1997; Huang & Louis, 1998; Strawderman, 2002) and censored medical costs (e.g. Lin *et al.*, 1997; Bang & Tsiatis, 2000; Huang & Lovato, 2002; Huang, 2002). Accumulation processes are non-decreasing stochastic processes and can be regarded as an integral,

$$\mathcal{Z}(t) = \int_0^t \mathcal{C}(u) du, \text{ where } \mathcal{C}(u) \text{ is the mark measured at time } u. \text{ In the analysis of lifetime}$$

medical costs, for example, $\mathcal{C}(u)$ is the cost at time u , T is death date and $\mathcal{Z}(T)$ is cumulative medical cost over the lifetime of the subject.

In this paper, we define the marked endpoint $Z = \mathcal{Z}(T)$ and ignore mark-scale information prior to time T , if it exists. Because our method makes no assumptions about the type of stochastic process that led to the mark, the methods may be applied widely to a diverse set of scientific problems. On the other hand, when $\mathcal{Z}(t)$ is an increasing stochastic process, as in the medical cost and quality-of-life examples before, ignoring information prior to T may be wasteful leading to inefficient estimation and less powerful statistical tests. For example, Bang & Tsiatis (2000) show how to construct locally efficient estimators for mean medical cost; see also, Strawderman (2002). The caveats of these better methods include modeling an obscure process with unique features that further rely on the disease process and, in some cases, the requirement that the marked process is indeed monotone. Moreover, Bang & Tsiatis (2000) showed that the improvement of the locally efficient estimators over the simple estimators can be modest.

In practice, one may not get to observe Z due to censoring. In this case, the observed stopping time is $X = T \wedge C$, where C is a random censoring time. Censoring on the time-scale introduces an identifiability concern on the mark-scale that complicates statistical methods for marginal mark-scale estimation and testing. Briefly, suppose Z is a real-valued marked end-point, so the random pair (T, Z) realizes values on $[0, \infty) \times \mathfrak{R}$. If the maximum support of C is less than the maximum support of T , then the marginal mark distribution is said to be nowhere identifiable (Lin *et al.*, 1997; Zhao & Tsiatis, 1997; Huang & Louis, 1998). Practically speaking, when the event time distribution has a long right-tail but the study has limited follow-up, then there is no opportunity to learn about the mark distribution corresponding to event times in the tail region, that is, for $(t, z) \in (\tau, \infty) \times \mathfrak{R}$ where τ is the maximum support of C . Thus, the potential non-identifiability of $G(z) = P(Z \leq z)$ in practice suggests that inference on the marginal mark distribution is infeasible.

Two strategies for drawing inference on marked endpoints yet overcoming the non-identifiability on its marginal distribution include redefining the endpoint and estimating the joint mark-and-time distribution. Let ξ be a user-defined, time-scale upper bound and define

the time-restricted random pair $(T^{\mathcal{E}}, Z^{\mathcal{E}})$, where $T^{\mathcal{E}} = T \wedge \xi$ and $Z^{\mathcal{E}} = \mathcal{Z}(T \wedge \xi)$. It is not difficult to show that the marginal distribution of the time-restricted marked endpoint $Z^{\mathcal{E}}$ is identifiable for $\xi \leq \tau$. Once the time-restricted endpoint $Z^{\mathcal{E}}$ is defined, one can estimate its moments or model the conditional expectation given covariates using methods of inverse weighting or other techniques for censored data (e.g. Lin *et al.*, 1997; Zhao & Tsiatis, 1997; Bang & Tsiatis, 2000; Strawderman, 2002). For a comprehensive review of the inverse weighting method with application to time-lagged endpoints and connections to monotone coarsening, see Tsiatis (2006).

An alternative strategy is to approach the marginal mark distribution $G(z)$ through the joint distribution of mark and time, that is,

$$G(z) = \int P(T \leq t, Z \leq z) = \int D(z|t)dF(t), \quad (1)$$

where $D(z|t) = P(Z \leq z | T = t)$, and $F(t) = P(T \leq t)$. The insight here is that both $D(z|t)$ and $F(t)$ in (1) are well defined and identifiable on $(-\infty, \tau] \times \mathfrak{R}$ in the presence of censoring. By manipulating the integrands on the right-hand side of (1) and integrating to the upper limit τ , one can define statistics that are both identifiable and address questions about the joint or marginal distributions (Huang & Louis, 1998; Huang & Lovato, 2002; Gilbert *et al.*, 2008). A key difference between the two strategies is that time-restricted mark $Z^{\mathcal{E}}$ is not the same endpoint as the one we initially stated was an outcome of interest. The scientific value of the time-restricted endpoint must be considered on a case by case basis, and, in many applications, it may be an adequate surrogate for the mark of primary interest. We eschew this particular issue by adopting a joint estimation framework in both time and mark-scales. A detailed review of the merits and disadvantages between inverse weighting and joint estimation for the analysis of marked endpoints is given elsewhere in the literature (Huang, 2009). Additional comments on limitations and future directions are relegated to the Discussion.

We propose two new families of nonparametric two-sample tests for testing the equality of marginal mark-scale distributions when marks are subject to induced censoring. The first family may be regarded as a nonparametric extension of a family of semi-parametric tests proposed by Huang & Lovato (2002). When the primary scientific interest lies in Z rather than $Z^{\mathcal{E}}$, there are few, if any, nonparametric two-sample tests for censored marks under general conditions. Gilbert *et al.* (2008) propose a suite of nonparametric tests for a different but related mark variable problem, and it is conceivable that some of their statistics could be reconfigured to test the composite null hypothesis in (2). It is worth noting that their test statistics require non-parametric smoothing of the time-varying relative hazard rate (Wang, 2005) between groups whereas the method proposed here requires no smoothing. Huang & Berry (2006) propose tests of the marginal mark distribution by modeling mark and time through bivariate copulas, but these tests, like the tests in Huang & Lovato (2002), are semi-parametric and require stronger modeling assumptions than what is proposed in the following sections.

The second family of tests are transformation-invariant tests of marked outcomes using a novel marked rank process. The proposed tests compare cumulative marked rank hazard functions between groups and have interesting operating characteristics because the marked outcome is missing when the corresponding failure time is censored. Developing transformation-invariant statistical tests is relevant substantively as the marginal mark distributions for some of the applications are known to be asymmetric with heavy tails. For example, the extremely long right-tail of the medical cost distribution has been well documented (e.g. Lin *et al.*, 1997; Bang & Tsiatis, 2000; Huang & Lovato, 2002; Huang, 2002). In practical work, it is common to take natural logarithms or other transformations of the marked outcome to remedy the sensitivity of data analyses to outlying observations. Unfortunately, currently available statistical tests (e.g. Huang & Lovato, 2002) are sensitive to such transformations and could potentially give contradictory conclusions in finite samples depending on the transformation. The proposed family of rank-based tests are invariant to monotone transformations of the marked outcome and more powerful than competing tests when $G(z)$ has heavy tails.

2. Notation and background

2.1. Notation

Let (T_k, Z_k) denote a random survival time and mark variable, respectively, from an arbitrary joint distribution for group k ($k = 1, 2$). In the sequel, our notation will be that (T_k, Z_k) realizes values in $\Re^2$, which deviates mildly from the range presented in Section 1. This range includes the possibility that time is transformed such that T_k can realize values on the entire real line. Modeling the natural logarithm of time is common in survival analysis (e.g. Kalbfleisch & Prentice, 2002, sec. 3.6, 7.1) and was used in Huang & Lovato (2002), for example. However, the logarithmic transformation is not a necessary condition of the methods proposed in this paper.

Now, consider the composite null hypothesis, 1

$$H_0: G_1(z) = G_2(z), \text{ for every } z, \quad G_k(z) = \int_{-\infty}^{\infty} D_k(z|t) dF_k(t), \quad F_k \text{ arbitrary}, \quad (2)$$

where $G_k(z) = P(Z_k \leq z)$, $D_k(z|t) = P(Z_k \leq z | T_k = t)$ and $F_k(t) = P(T_k \leq t)$. Our objective is to construct statistical tests of (2) with censored observations, $(X_k, \delta_k, \delta_k Z_k)$, $X_k = T_k \wedge C_k$, $\delta_k = I(T_k \leq C_k)$, C_k is a censoring random variable assumed to be independent of T_k and Z_k and whose distribution may differ by group.

As we argued in Section 1, statistical tests of the composite null hypothesis in (2) can be motivated by expression (1). Let $\psi(\cdot)$ be any monotone function. Then, the cumulative marked hazard function is defined as

$$\int_{-\infty}^{\infty} \int_{-\infty}^t \psi(z) h_k^{\#}(u, z) du dz = \int_{-\infty}^t E\{\psi(Z_{k1}) | T_{k1} = u\} h_k(u) du, \quad (3)$$

for $h_k(t) = \lim_{\varepsilon \rightarrow 0} \varepsilon^{-1} P(t \leq T_{k1} < t + \varepsilon | T_{k1} \geq t)$, and $h_k^\#(t, z)$ is the mark-specific hazard rate function (Huang & Louis, 1998),

$$h_k^\#(t, z) = \lim_{\varepsilon_1, \varepsilon_2 \rightarrow 0} (\varepsilon_1, \varepsilon_2)^{-1} P(t \leq T_{k1} < t + \varepsilon_1, z \leq Z_{k1} < z + \varepsilon_2 | T_{k1} \geq t).$$

Note that the expression on the right-hand side of (3) is exactly in the form of right-hand of (1) except the integrator is $H_k(t) = \int_{-\infty}^t h_k(u) du$, a 1-1 transformation of $F_k(t)$.

In our two-sample setup, the observed data are $(X_{ki}, \delta_{ki}, \delta_{ki} Z_{ki})$, $i = 1, \dots, n_k$, assumed to be mutually independent and follow a common distribution within group, say, the distribution of $(X_{11}, \delta_{11}, \delta_{11} Z_{11})$ or $(X_{21}, \delta_{21}, \delta_{21} Z_{21})$ for group 1 or 2, respectively. If we define the marked process $I(X_{ki} \leq t, \delta_{ki} = 1) \psi(Z_{ki})$ where $I(\cdot)$ is the indicator function, then a uniformly and strongly consistent for is estimator the right-hand side of (3) is given by

$n_k^{-1} \sum_{i=1}^{n_k} \int_{-\infty}^t \psi(Z_{ki}) d\{I(X_{ki} \leq u, \delta_{ki} = 1)\} / Y_{Nk}(u)$, where $Y_{Nk}(t) = n_k^{-1} \sum_{i=1}^{n_k} I(X_{ki} \geq t)$. This statistic was first proposed and investigated by Huang & Louis (1998) and is the nonparametric maximum likelihood estimator for the cumulative marked hazard function in (3) when $\psi(\cdot)$ is fixed and known.

2.2. Existing semi-parametric tests

Rather than allowing arbitrary time-scale distributions F_k as in (2), suppose we assume the location-shift model

$$F_1(t) = F_2(t - \theta), \quad (4)$$

where $\omega(t)$ is an unknown parameter. For some applications, the location-shift model in (4) is sensible on a properly transformed time-scale. For example, a location change on the logarithmic scale implies a scale change on the natural scale of the event time. In Huang & Lovato (2002), time is modeled on the natural logarithmic scale.

Next, set $\psi(\cdot) = \cdot$ in (3) and define the *marked identity process*, $I(X_{ki} \leq t, \delta_{ki} = 1) Z_{ki}$ and the empirical average of marked identity processes, $A_{Nk}(t) = n_k^{-1} \sum_{i=1}^{n_k} Z_{ki} I(X_{ki} \leq t) \delta_{ki}$. Then, the cumulative difference in marked hazard functions is

$$\int_{-\infty}^{\tau} \omega(t) \left[\frac{dA_{N1}(t)}{Y_{N1}(t)} - \frac{dA_{N2}(t - \theta)}{Y_{N2}(t - \theta)} \right], \quad (5)$$

where $\omega(t)$ is a weight function. As noted previously, $\int_{-\infty}^{\tau} dA_{N1}(t)/Y_{N1}(t)$ is the maximum likelihood estimator of the cumulative marked hazard function, $\int_{-\infty}^{\tau} E(Z_{11}|T_{11}=t)h_1(t)dt$, and, consequently, (5) is the maximum likelihood estimator of the weighted difference in cumulative marked hazard functions. In practice, θ must be estimated from the data; let $\hat{\theta}_N$ be an estimator for θ . After we replace θ with $\hat{\theta}_N$ in (5), then (5) may be used to test the null hypothesis in (2) under the location-shift model in (4). To account for the uncertainty in $\hat{\theta}_N$, Huang & Lovato (2002) proposed a minimum dispersion test built around a pair of estimating functions. Expression (5) is the first estimating function because it has mean zero under the null hypothesis while the second estimating function is due to estimating the nuisance parameter θ . We refer the interested reader to the literature for further details of their procedure.

In the analysis of event-time data, two common weight functions are the Gehan and log-rank weight function given by $\omega(t) = Y_{N1}(t)Y_{N2}(t)$ and $\omega(t) = Y_{N1}(t)Y_{N2}(t)/Y_{Nc}(t)$, respectively, where $Y_{Nc}(t) = N^{-1}\{n_1 Y_{N1}(t) + n_2 Y_{N2}(t)\}$ and $N = n_1 + n_2$. Under the local sequence of alternative hypotheses $N^{1/2}\{\mu_1(t) - \mu_2(t - \theta)\} \rightarrow \gamma(t)$ uniformly on $t \in \mathfrak{R}$ for some bounded, deterministic function $\gamma(t)$ and arbitrary θ in (4) where $\mu_k(t) = E(Z_{k1}|T_{k1} = t)$, Huang & Lovato (2002) showed that the log-rank weight function was locally optimal if the mark and survival time were independent. We restrict our attention to the Gehan and log-rank weight functions to facilitate comparisons with earlier work and because they are well known in the survival analysis literature (e.g. Fleming & Harrington, 1990, ch. 7) and similar to ones proposed and considered by Gilbert *et al.* (2008, 3.2) for a related problem.

3. New methods

3.1. Nonparametrically calibrated linear tests

Although the semi-parametric model (4) may hold in certain cases, it does not hold in general. When the underlying assumptions of the semi-parametric test derived from (5) fail, one can have statistical tests with suboptimal operating characteristics. For these reasons, it is desirable to relax this assumption.

To define the nonparametric test, we need some additional notation. Let $S_k(t)$ be the marginal survivor function, $S_k(t) = 1 - F_k(t)$, and $S_k^{-1}(p) = \inf_t \{t | S_k(t) \leq 1 - p\}$. Suppose that $S_{Nk}(t)$ is a uniformly consistent estimator of the marginal survivor function $S_k(t)$, for example, the Kaplan-Meier or Nelson-Aalen estimator. Then, for the marked identity process in (5), a nonparametric estimate of the weighted difference in cumulative hazard functions is

$$L = \int_0^{p_{\max}} \omega(p) \left[\frac{dA_{N1}\{S_{N1}^{-1}(p)\}}{Y_{N1}\{S_{N1}^{-1}(p)\}} - \frac{dA_{N2}\{S_{N2}^{-1}(p)\}}{Y_{N2}\{S_{N2}^{-1}(p)\}} \right], \quad (6)$$

where $\omega(p)$ is the fractile-analogue of the weight function $\omega(t)$ in (5) and $p_{\max} - 1$ is an upper limit on the fractile process to ensure identifiability, again, analogous to the role that τ played in (5). Both L in (6) and the expression in (5) may be viewed as extensions of weighted log-rank statistics (Fleming & Harrington, 1990) to marked processes. In order to use the statistic L to test the null hypothesis in (2), we need to know its distribution under the null hypothesis; we defer this discussion to Section 3.4.

Finally, to see how our proposed statistic L relates to the semi-parametric test discussed in Section 2.2, let $t = S_{N1}^{-1}(p)$ in (6) and rewrite L as

$$(6) = \int_{-\infty}^{\tau} \omega(t) \left[\frac{dA_{N1}(t)}{Y_{N1}(t)} - \frac{dA_{N2}\{t - \hat{\theta}_N(t)\}}{Y_{N2}\{t - \hat{\theta}_N(t)\}} \right], \quad (7)$$

where $\hat{\theta}_N(t) = t - S_{N2}^{-1}\{S_{N1}(t)\}$ is an estimator of $\theta(t) = t - S_2^{-1}\{S_1(t)\}$ and $\tau = S_1^{-1}(p_{\max})$. If we assume $\theta(t) \equiv \theta$ for all t , $t \leq \tau$, and the common location parameter θ was estimated with the weighted log-rank estimating function (Tsiatis, 1990), then (7) reduces to (5) proposed by Huang & Lovato (2002).

3.2. The marked rank process and rank-based tests

To improve the robustness of L , it is desirable to introduce a function $\psi(\cdot)$ in (3) that limits the influence of extremely large marks. In practice, the ranks of the observations moderate the influence of extreme marks and are invariant to transformations of the marked outcomes. However, the true rank among all marks (Z_{ki} , $i = 1, \dots, n_k$, $k = 1, 2$) is not available as the mark is missing when the corresponding event time is censored. So we propose a novel marked rank process and derive the induced expression for $\psi(\cdot)$ from the definition of cumulative marked rank hazard function.

Our new marked process is conceived by taking the marked identity process, $I(X_{ki} \leq t, \delta_{ki} = 1)Z_{ki}$ in Section 2.2 and replacing the mark Z_{ki} with its rank, using the observable data. Let $\mathcal{R}_{2,1}^u(Z_h)$ be the rank of Z_h among the uncensored marks in both samples and define the marked rank process for the i th subject in group k , $N^{-1}I(X_{ki} \leq t, \delta_{ki} = 1)\mathcal{R}_{2,1}^u(Z_{ki})$. Then, we define of average of marked rank processes

$$\begin{aligned} Q_{Nk}(t) &= (Nn_k)^{-1} \sum_{i=1}^{n_k} \mathcal{R}_{2,1}^u(Z_{ki}) I(X_{ki} \leq t) \delta_{ki}, \quad (8) \\ &= \int_{-\infty}^{\infty} \left\{ \frac{n_1}{N} C_{N1}(z) + \frac{n_2}{N} C_{N2}(z) \right\} B_{Nk}(t, dz), \end{aligned}$$

where $B_{Nk}(t, z)$ and $C_{Nk}(z)$ are the empirical processes,

$$B_{Nk}(t, z) = \frac{1}{n_k} \sum_{i=1}^{n_k} I(X_{ki} \leq t, Z_{ki} \leq z) \delta_{ki}, \quad C_{Nk}(z) = \frac{1}{n_k} \sum_{i=1}^{n_k} I(Z_{ki} \leq z) \delta_{ki}.$$

Using these definitions, we propose the statistic

$$U_1 = \int_0^{p_{\max}} \omega(p) \left[\frac{dQ_{N1}\{S_{N1}^{-1}(p)\}}{Y_{N1}\{S_{N1}^{-1}(p)\}} - \frac{dQ_{N2}\{S_{N2}^{-1}(p)\}}{Y_{N2}\{S_{N2}^{-1}(p)\}} \right].$$

Similar to L in (6), U_1 is the cumulative difference in marked hazard rates between two groups. But, unlike L , U_1 is invariant to monotone transformations of the marks and robust to mark distributions with heavy tails. Under noninformative censoring, one can show that cumulative marked rank hazard function is (3) with $\psi(z) = \{n_1 W_1(z) + n_2 W_2(z)\}/N$, where

$$W_k(z) = \int_{-\infty}^{\infty} \int_{-\infty}^z S_k^o(r) \bar{f}_k(r, v) dv dr, \quad \bar{f}_k(t, z) \text{ is the joint probability density function for the}$$

random pair (T_{k1}, Z_{k1}) and $S_k^o(t) = P(C_{k1} > t)$. Thus, the cumulative marked rank hazard function in each group depends on the entire distribution of $(T_{11}, Z_{11}, C_{11}, T_{21}, Z_{21}, C_{21})$, in contrast to the cumulative marked identity hazard function, which only depends on the distribution of (T_{11}, Z_{11}) .

To have some sense of the alternative hypothesis using the identity and marked rank processes, it is constructive to consider the location-shift model as in Huang & Lovato (2002); for example, see the statistic L in (7) with $\theta(t)$ for all $t \leq \tau$. Under this restriction, our alternative hypothesis is that $E\{\psi(Z_{11})|T_{11} = t\} = E\{\psi(Z_{21})|T_{21} = t - \theta\}$ for some $t \leq \tau$, note $\psi(\cdot)$ is a type of probability integral transformation of the marked endpoint. On the other hand, the alternative hypothesis for the marked identity process is $E(Z_{11}|T_{11} = t) = E(Z_{21}|T_{21} = t - \theta)$ for some $t \leq \tau$. Thus, linear tests rejects the null hypothesis if the cumulative differences in conditional expectations of the marked end-points given time are sufficiently large whereas rank-based tests reject the null hypothesis if the cumulative difference in conditional probabilities is sufficiently large.

3.3. Marked weighted rank processes

Although U_1 has desirable robustness properties, it does depend on the distribution of the censoring random variables C_{11} and C_{21} . Because censoring distributions are viewed as nuisance parameters in survival analysis, it is preferable that our tests not depend on these distributions. To overcome this potential deficiency, we introduce a marked weighted rank process.

Let $\phi_{ki} = \{S_k^o(X_{ki})\}^{-1}$, $S_k^o(t) = P(C_{k1} > t)$ as defined in Section 3.2, and let $\mathcal{R}_{2,\phi}^u(Z_h)$ be the weighted rank of Z_h among the uncensored marks in both samples, that is,

$$\mathcal{R}_{2,\phi}^u(Z_h) = \sum_{i=1}^{n_1} I(Z_{1i} \leq Z_h) \delta_{1i} \phi_{1i} + \sum_{j=1}^{n_2} I(Z_{2j} \leq Z_h) \delta_{2j} \phi_{2j}.$$

Then, define the average of marked weighted rank processes,

$$Q_{Nk}(t; \phi) = \frac{1}{N n_k} \sum_{i=1}^{n_k} \mathcal{R}_{2,\phi}^u(Z_{ki}) I(X_{ki} \leq t) \delta_{ki}, \quad (9)$$

and the weighted cumulative difference in marked weighted rank hazard rates,

$$U_\phi = \int_0^{p_{\max}} \omega(p) \left[\frac{dQ_{N1}\{S_{N1}^{-1}(p); \phi\}}{Y_{N1}\{S_{N1}^{-1}(p)\}} - \frac{dQ_{N2}\{S_{N2}^{-1}(p); \phi\}}{Y_{N2}\{S_{N2}^{-1}(p)\}} \right]. \quad (10)$$

The difference between the weights $\omega(p)$ and ϕ_{ki} is that the former weights time-scale differences in the group-specific marked hazard functions whereas ϕ_{ki} is a mark-scale subject-specific weight in the measure $Q_{Nk}(t; \phi)$. Clearly, when $\phi_{ki} \equiv 1$, then $\mathcal{R}_{2,\phi}^u(Z_h)$, $Q_{Nk}(t; \phi)$ and U_ϕ reduce to $\mathcal{R}_{2,\phi}^u(Z_h)$, $Q_{Nk}(t)$ and U_1 , respectively.

Using this definition of weight ϕ_{ki} and under non-informative censoring, one can show that the corresponding $\psi(\cdot)$ in (3) is $\psi(z) = \{n_1 G_1(z) + n_2 G_2(z)\}/N$, where $G_k(z)$ is the marginal distribution of Z_{k1} . Hence, the cumulative marked weighted rank hazard function depends on the distribution of $(T_{11}, Z_{11}, T_{21}, Z_{21})$. Although the survivor distribution $S_k^o(t)$ in ϕ_{ki} is unknown, the Kaplan–Meier estimator $S_{Nk}^o(t)$ is a uniformly and strongly consistent estimator for $S_k^o(t)$. In the sequel, the statistic U_ϕ is defined with estimated weights $\{S_{Nk}^o(X_{ki})\}^{-1}$.

Remark 1—It is interesting to note that $\phi_{ki} = \{S_{Nk}^o(X_{ki})\}^{-1}$ is the conventional weight used in inverse weighting methodology applied to censored data problems (e.g. Tsiatis, 2006). Authors have extended this technique to one and two-sample problems with marked endpoints and time-lagged responses (e.g. Zhao & Tsiatis, 1997; Bang & Tsiatis, 2000). One noteworthy difference in adopting the technique here is that when $\phi_{ki} \equiv 1$, the statistic U_ϕ reduces to U_1 and remains a valid statistic for testing (2). On the other hand, if one sets $\phi_{ki} \equiv 1$ in statistics built on a theory of inverse probability of censoring, the statistic reduces to one based on the uncensored observations and typically has limited scientific value. For example under suitable regularity conditions, $n_k^{-1} \sum_{i=1}^{n_k} \delta_{ki} I(X_{ki} > t) \phi_{ki}$ is a consistent estimator of $S_k(t)$ whereas $n_k^{-1} \sum_{i=1}^{n_k} \delta_{ki} I(X_{ki} > t)$ is a consistent estimator of $P(T_{k1} > t, T_{k1} < C_{k1})$.

Remark 2—In this paper, we consider one definition of the weight ϕ_{ki} that induces certain properties on the test statistic U_ϕ in (10). However, the definition of marked weighted rank processes in (9) is general, and other weights could be considered. The operating characteristics of test statistics U_ϕ using different weights ϕ_{ki} would need to be considered on a case-by-cases basis.

3.4. Practical notes

We construct asymptotically valid critical regions based on the large sample distribution of L , U_1 and U_ϕ . In the Appendix, we argue that L , U_1 and U_ϕ converge in distribution to normal random variables with finite variance. Then, for example, a rank-based level- α two-sided test rejects the null hypothesis for $|U_1/\{\text{var}(U_1)\}^{1/2}| > z_{1-\alpha/2}$, where $z_{1-\alpha/2}$ are critical values from the standard normal distribution. Although none of $\text{var}(L)$, $\text{var}(U_1)$, or $\text{var}(U_\phi)$ can be evaluated directly, they can be estimated consistently through resampling (Kosorok, 2008) as we now show.

Let $\{\zeta_{ki}, i = 1, \dots, n_k, k = 1, 2\}$ be independent and identically distributed random variables with unit mean and variance and completely independent of the observed data. Define the following empirical processes:

$$Y_{Nk}^*(t) = \frac{1}{n_k} \sum_{i=1}^{n_k} \zeta_{ki} I(X_{ki} \geq t), \quad A_{Nk}^*(t) = \frac{1}{n_k} \sum_{i=1}^{n_k} \zeta_{ki} Z_{ki} I(X_{ki} \leq t) \delta_{ki},$$

$$C_{Nk}^*(z; \phi) = \frac{1}{n_k} \sum_{i=1}^{n_k} \zeta_{ki} I(Z_{ki} \leq z) \delta_{ki} \phi_{ki}^*, \quad B_{Nk}^*(t, z) = \frac{1}{n_k} \sum_{i=1}^{n_k} \zeta_{ki} I(X_{ki} \leq t, Z_{ki} \leq z) \delta_{ki},$$

where the weights ϕ_{ki}^* are estimated from the perturbed Kaplan–Meier or Nelson–Aalen estimator, if necessary. For example, the perturbed Nelson–Aalen estimator of $S_k^o(t)$ is $S_{Nk}^{o*}(t) = \exp\{-H_{Nk}^{o*}(t)\}$, where

$$H_{Nk}^{o*}(t) = \int_{-\infty}^t \{Y_{Nk}^*(u)\}^{-1} dN_{Nk}^{o*}(u), \quad N_{Nk}^{o*}(t) = n_k^{-1} \sum_{i=1}^{n_k} \zeta_{ki} I(X_{ki} \leq t, \delta_{ki} = 0).$$

Then, the resampled statistics L^* , U_ϕ^* are given by

$$L^* = \int_0^{p_{\max}} \omega^*(p) \left[\frac{dA_{N1}^*\{S_{N1}^{-1}(p)\}}{Y_{N1}^*\{S_{N1}^{-1}(p)\}} - \frac{dA_{N2}^*\{S_{N2}^{-1}(p)\}}{Y_{N2}^*\{S_{N2}^{-1}(p)\}} \right],$$

$$U_\phi^* = \int_0^{p_{\max}} \omega^*(p) \left[\frac{dQ_{N1}^*\{S_{N1}^{-1}(p); \phi\}}{Y_{N1}^*\{S_{N1}^{-1}(p)\}} - \frac{dQ_{N2}^*\{S_{N2}^{-1}(p); \phi\}}{Y_{N2}^*\{S_{N2}^{-1}(p)\}} \right],$$

where $\omega^*(p)$ is the perturbed weight function defined through $Y_{Nk}^* = \{S_{Nk}^{-1}(p)\}$ and the perturbed marked rank empirical measure is

$$Q_{Nk}^*(t; \phi) = \int_{-\infty}^{\infty} \left\{ \frac{n_1}{N} C_{N1}^*(z; \phi) + \frac{n_2}{N} C_{N2}^*(z; \phi) \right\} B_{Nk}^*(t, dz).$$

Under the conditions given in the Appendix, conditional on the observed data.

$N^{1/2}(U_{\phi}^* - U_{\phi})$ has the same asymptotic distribution as $N^{1/2}\{U_{\phi} - E(U_{\phi})\}$, and, hence, the sample variance of the normalized differences, $(U_{\phi}^* - U_{\phi})$, is a consistent estimator the sampling variance of the statistic, U_{ϕ} . Thus, we repeatedly generate perturbations $\{\zeta_{ki}, i = 1, \dots, n_k, k = 2\}$ a large number of times, say M times, and compute the test statistic, $U_{\phi}/\{\text{var}(U_{\phi})\}^{1/2}$, using the sample variance of $\{(U_{\phi, m}^* - U_{\phi}), m = 1, \dots, M\}$ in place of $\text{var}(U_{\phi})$ in the denominator. We compare observed test statistic to critical values from the normal distribution for a level- α test. The resampling plans for L and U_1 follow a similar outline. Other variations and refinements of statistical tests are discussed briefly in Section 6.

4. Simulation studies

We evaluated the size and power of the proposed tests through simulation studies. To facilitate a comparison of nonparametric tests with semi-parametric tests, our simulation design follows a strategy similar to Huang & Lovato (2002). We assume the location-shift model, $(T_2, Z_2) = (T_1 + \theta, Z_1 + \eta)$, with $\theta = 0.2$, (T_1, Z_1) follows Frank's bivariate copula, and the marginal time-scale distribution $F_k(t)$ is standard extreme value. We adopt Genest's (1987, p.554) formulation of Frank's bivariate copula through the association parameter γ ; the values are $\gamma = 0$ or 10 and interpreted as Kendall's rank correlation, $\tau = 0$ or 0.67, respectively. The censoring distributions of $\exp(C_k)$ were $\text{Unif}(\theta, \kappa)$, where κ is 5.2 or 2.4 resulting in approximately 20% and 40% censored observations, respectively. The marginal distribution $G_1(z)$ is either standard uniform or standard exponential. In the case where $G_1(z)$ is a standard uniform distribution, the simulation setup here is identical to the setup in Huang & Lovato (2002). A representative subset of the simulation results are presented in Figs 1, 2 and 3 whereas the complete tables of simulation results are available in Online Supporting Information. The size of the two-sample tests at the nominal level with the Gehan and log-rank weight function is summarized in Fig. 1 for the exponential marginal mark distribution, and the statistical power is summarized in Figs 2 and 3 for both uniform and exponential mark distributions. In each figure, the organization of six subfigures is as follows:

- simulation results for complete-case naïve tests are in panels (a) and (d); results for proposed tests with Gehan weight are summarized in panels (b) and (e); results for proposed tests with log-rank weight are summarized in panels (c) and (f);
- simulation results in panels (a), (b) and (c) are summarized across four sample sizes, $n_k = 25, 50, 100$ and 200 with the censoring proportion fixed at 20%;
- in panels (d), (e) and (f), simulation results are displayed across a range of censoring proportions 0%, 20% and 40% with Kendall's τ fixed at $\tau = 0.67$.

Finally, each panel of each figure has two legends that describe the precise meaning of the different line colours and textures used in each display.

We begin our summary of the numerical results with the empirical size of the two-sample tests in Fig. 1. In panels (a) and (d), we observe that both the two-sample t -test and the Wilcoxon–Mann–Whitney (WMW) tests have the correct size when there is no association between mark and time, that is, $\tau = 0$, or when the censoring proportion is zero. However, in the presence of non-zero mark-time association and non-zero censoring proportion, these tests have the incorrect size. This conclusion is not an artifact of the heavy tail in the mark distribution as both the t -test and robust WMW test are similarly affected, and the same result is true for uniform mark distributions. As we can see in panel (d), this conclusion is not a small sample problem because the size of tests gets worse as the sample size increases. In particular, when $n_k \geq 200$ and the censoring rate is 40%, the empirical size of the two-sample t -test is about 12.5% and 10.9% for the WMW test, more than twice the nominal level (result not shown). In panels (b), (c), (e) and (f), we observe that the empirical sizes of the proposed tests are close to nominal level. Additional comments regarding deviation from the nominal size for our proposed tests and possible finite sample improvements are relegated to the Discussion in Section 6.

The empirical power of the two-sample tests under location-shift alternative hypotheses is summarized in Figs 2 and 3. When the marginal mark distribution is uniform in Fig. 2, we find that the linear tests L are slightly more powerful than the rank-based tests, uniformly across sample sizes, censoring proportions and weight functions. The more powerful of the rank-based tests with Gehan or log-rank weight depends on the censoring proportion and association parameter τ . In the presence of censoring, tests with Gehan weight were more powerful than tests with log-rank weight when T_k and Z_k were correlated and vice versa when T_k and Z_k were independent, the latter of which agrees heuristically with the local optimality of the log-rank weight function in Huang & Lovato (2002, Appendix). We found that U_1 and U_ϕ had similar statistical power in the simulation scenarios considered here. Compared with the semi-parametric tests proposed in Huang & Lovato (2002), the nonparametric tests were no worse than the semi-parametric tests in this simulation design and even appeared to be mildly better.

When the marginal mark distribution is standard exponential in Fig. 3, the rank-based tests are approximately 40% more powerful than the linear tests, on average, uniformly across sample size, weight function and censoring proportion. In this simulation scenario, we also observe a striking difference in statistical power among the proposed tests in panels (a)–(c) when $\tau = 0$ versus $\tau = 0.67$. For example, when the sample size is $n_k = 100$, the statistical power in the linear Gehan test drops about 50% from 0.8 to 0.3 when $\tau = 0$ compared with when $\tau = 0.67$. The analogous log-rank linear test drops only about 10% statistical power from 0.47 to 0.36. In this scenario and in contrast to what we observed in Fig. 2, we found the performance of U_1 to be somewhat better than U_ϕ , where the improvement depended on censoring proportion, sample size, Kendall's τ and weight function $\omega(p)$. Although the complete-case t -test and WMW test appear to have statistical power similar or in excess of the proposed statistical tests, the conclusion is misleading because the naïve tests are not

valid tests in the presence of both a non-zero censoring rate and mark-time association, in general.

5. Data analysis

To illustrate the utility of the two-sample tests in practice, we re-analysed data from a study of non-small cell lung cancer. Southwest Oncology Group Study 9509 was a randomized clinical trial of Vinorelbine plus Cisplatin versus Paclitaxel plus Carboplatin among untreated participants with advanced stage lung cancer. The study collected clinical measurements as well as resource utilization measures during the follow-up period at the 3, 6, 12, 18 and 24-month visits. Lifetime medical cost is the sum of costs associated with resources utilized from the date of randomization until the death date. Medical costs were standardized to 1998 US dollars according following guidelines in the medical care component of the Consumer Price Index. The data include medical cost and survival data on $N = 398$ participants, 198 of whom were randomized to Vinorelbine plus Cisplatin. The overall censoring rate was 36% and median survival time was 6 months. The same data set was previously analysed by Huang & Lovato (2002) using their family of semi-parametric tests.

Figure 4 presents group-specific comparisons of survival and cumulative marked hazard rate functions. Panel (a) displays Kaplan–Meier estimates of the group-specific survival distributions and suggests that there is basically no difference. A formal log-rank test for the equality of two survivor distributions fails to reject the null hypothesis at the nominal level (log-rank test statistic = 0.20, p -value = 0.655) and confirms this conjecture; this point was previously reported by the SWOG 9509 study investigators. We also investigated group-specific censoring rates, and, like the survivor distributions, censoring distributions are quite similar (figure not shown). A formal test of the equality of group-specific censoring distributions also fails to reject the null hypothesis at the nominal level (log-rank test statistic = 0.03, p -value = 0.868).

The cumulative marked hazard functions are displayed in panels (b)–(d). The cumulative marked identity hazard functions, $A_{Nk}(t)$, $k = 1, 2$, in panel (b) are displayed on the natural logarithm of medical cost whereas Huang & Lovato (2002, Fig. 1) present estimates on the original scale. In contrast to the cumulative marked identity process, the cumulative marked rank and weighted rank functions, $Q_{Nk}(t)$ and $Q_{Nk}(t, \phi)$ in panels (c) and (d), respectively, are invariant to monotone transformations of the mark and do not change whether we are estimating the functions on the original or logarithmic scale. Visually, all three panels (b)–(d) seem to suggest some amount of cost-scale differences between the two treatment groups. Note the magnitude on the y -axis is very different in panel (b) compared with panels (c)–(d). Heuristically and informally, the cost-scale differences for the rank-based marked hazard functions appear somewhat larger than those based on the cumulative marked identity hazard rate processes and functions.

We computed the proposed test statistics and compared them with normal critical values to formally test the equality of group-specific marginal cost distributions as given in (2). Recall, our test statistics are defined as one of L , U_1 or U_ϕ (Numerator) divided by its

standard deviation. In Table 1, we present L , U_1 and U_ϕ for both Gehan and log-rank weight functions, the Monte Carlo standard deviation (SD) of the resamples, the median absolute deviation from the mean divided by 0.6745 (Rob.SD), the numerator divided by the standard deviation is the z -score and the p -value is computed as the $2\{1 - \Phi(|z|)\}$, $\Phi(\cdot)$ is the Gaussian cumulative distribution function and the z is the z -score.

In Table 1, when we model log-transformed medical cost, we see that the test statistics are all quite similar. The log-rank weight function leads to larger numerator in absolute value relative to the Gehan weight function for each of L , U_1 and U_ϕ but also more variability, which ultimately leads to a smaller test statistic (in absolute value). Consequently, the p -values for the log-rank statistic are also higher, although still statistically significant at the nominal level. We also find that both rank-based tests led to very similar results, regardless of whether weights were included or not. In the Vinorelbine plus Cisplatin group, the range of the estimated weights $\hat{\phi}_{ki}$ was 1.00–2.28, and the inner quartile range was 1.06–1.35; in the Paclitaxel plus Carboplatin group, the range was 1.00–2.17, and the inner quartile range was 1.08–1.30. We see that the numerator U_ϕ was larger than numerator U_1 , which agrees with panels (c)–(d) in Fig. 4. But the overall test statistic for U_ϕ is smaller than U_1 because the variance of U_ϕ is larger than U_1 because of the estimation of the survivor function $S_k^o(t)$. Thus, the conclusion from the statistical tests is that Paclitaxel plus Carboplatin results in significantly higher lifetime medical cost compared to Vinorelbine plus Cisplatin.

Interestingly, the absolute value of the test statistics based on L decreases substantially when we consider medical cost on its natural scale. If we considered more stringent level-0.01 statistical tests, we would no longer reject the null hypothesis using L . However, rank-based tests are invariant to such transformations, and the fact that mark-scale transformations can and do make a difference in practice was one of main theses of this paper. Under a different set of circumstances, the change in statistical significance of the two-sample test based on L at the 0.01-level could instead be a change in statistical significance at the nominal level and could impact profoundly the enthusiasm and support for an experimental treatment.

Finally, our tests rely on asymptotic normality to define critical regions that have approximately the correct size for N sufficiently large. Imprecise estimation of the sampling variance of L , U_1 and U_ϕ or non-normal sampling distributions could weaken the applicability of the proposed tests. To have some sense of the sampling distribution of the test statistics, we examine in Fig. 5 histograms of the perturbed test statistics $(L^* - L)$, $(U_1^* - U_1)$, and $(U_\phi^* - U_\phi)$ from 1000 resamples. Evidently, the histograms suggest the sampling distribution is approximately Gaussian. We took the additional step of examining quantile–quantile plots and then formally testing the empirical percentiles of the resamples against percentiles of a Gaussian random variable. The range of the p -values from six Anderson–Darling tests corresponding to the six histograms of resamples in Fig. 5 was 0.35–0.80; hence, there is no evidence to reject the hypothesis that resamples are normally distributed. Thus, in the current data analysis, the sampling distribution for the test statistics appeared to be Gaussian and consistent with a presumption of our statistical tests.

6. Discussion

We propose two families of non-parametric two-sample tests for the composite null hypothesis that marginal mark are equal given arbitrary time-scale distributions. One family of tests is a non-parametric extension of a family of semi-parametric tests proposed by Huang & Lovato (2002); a second family of tests is based on a novel marked rank process. The latter methods were motivated by the fact that the distributions for some marked endpoints of interest, for example, medical cost, are skewed, heavy-tailed or both. In such cases, robust statistical procedures will tend to provide a better summary of two-sample differences. For example, we found that rank-based tests could be 40% more powerful than linear tests when marked endpoints had a standard exponential distribution. Moreover, the rank-based tests are invariant to monotone transformations of the marked endpoint while the existing linear tests are not.

The general strategy of the two-sample tests is to test for differences in the cumulative marked hazard functions on the support of the failure time distribution; these cumulative differences were L , U_1 or U_ϕ defined through marked processes $A_{N\kappa}(t)$, $Q_{N\kappa}(t)$ or $Q_{N\kappa}(t)$, respectively, given in Section 3. The test statistics were defined as one of L , U_1 or U_ϕ divided by an estimate of its variance and argued such statistic ought to have a Gaussian distribution for sufficiently large N . We estimated the sampling variance by repeatedly perturbing the data and taking the sample variance of the resampled statistics; this method worked well for our analysis of the data from SWOG 9509 in Section 5. In small samples, we expect that a large sample perspective could be inadequate and better testing procedures desirable. These better procedures may include bootstrap or permutation tests, and we leave this for future work.

A competing statistical and philosophical framework to the one taken here is to define a time-restricted marked endpoint $Z^\mathcal{E}$ and adopt inverse probability weighting to estimate its marginal distribution. As we stated in Section 1, restricting the marked endpoint addresses the potential non-identifiability of the marginal mark distribution directly, so identifiability is not a concern. Another advantage of inverse weighting is that the methods are relatively mature by now and straightforward to implement. A caveat of the time-restricted marked endpoint is that its marginal distribution can be very different than the distribution of the original, unrestricted marked endpoint; for example, see Huang (2009). Thus, conclusions drawn from the analysis of time-restricted marked endpoints could be different than conclusions drawn from analyses of the unrestricted marked endpoints. So a data analyst's decision to redefine a marked endpoint and draw inference on the time-restricted scale should be taken seriously and its potential consequences discussed with the investigator. Alternatively, one can use methods that evaluate the original marked endpoint, like the methods in Huang & Lovato (2002), Gilbert *et al.* (2004), Gilbert *et al.* (2008) or the two-sample tests proposed here.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

We thank the associate editor and two anonymous referees for comments that improved the manuscript. We gratefully acknowledge Drs John Crowley and Mary Redman for sharing the SWOG data and thank Yijian Huang for reading earlier drafts of this manuscript. This work was supported in part by US NIH Grant P30AI078498.

Appendix

Let T_k , C_k and Z_k be continuous random variables defined over the real line $\mathfrak{R}$. Using notation introduced earlier, assuming that C_k is independent of (T_k, Z_k) , the absolutely continuous distribution function of (T_k, C_k, Z_k) is $\bar{F}_k(t, z)F_k^O(c)$ with density function $\bar{f}_k(t, z)f_k^O(c)$. In addition, we impose the following two regularity conditions.

Condition A1

There exists $\tau_l, \tau_b, \tau_u, \tau_r$ such that $\bar{f}_k(t, z)S_k^O(t) > 0$ for every $(t, z) \in [\tau_b, \tau_u] \times [\tau_l, \tau_r]$ where $S_k^O(t) = 1 - F_k^O(t)$.

Condition A2

The weight function $\omega(t)$ converges to a deterministic limit $\dot{\omega}(t)$, uniformly in t , $t \in [\tau_b, \tau_u]$.

Condition A1 allows us to ensure the identifiability of the fractile process, and Condition A2 is a standard condition placed on the weight function in the analysis of the weighted log-rank estimating function. Both the Gehan and log-rank weight function satisfy this latter condition.

When we first began investigating the asymptotic properties of the proposed two-sample tests, we had adopted the notation in (7) that more closely mimicked Huang & Lovato (2002) before we rewrote the integrals as in (6). Without loss of generality, we retain this notation through the following technical results.

Proposition 1

Assuming that C_k is independent of (T_k, Z_k) , the expectations of L , U_1 and U_ϕ are

$$\begin{aligned} E(L) &= \int_{-\infty}^{\tau} \dot{\omega}(t) \left[E(Z_{11} | T_{11} = s) h_1(s) \right. \\ &\quad \left. - E\{Z_{21} | T_{21} = s - \theta(s)\} h_2\{s - \theta(s)\} \right] ds, \\ E(U_1) &= \int_{-\infty}^{\tau} \dot{\omega}(t) \left[E\left\{ \frac{n_1}{N} W_1(Z_{11}) + \frac{n_2}{N} W_2(Z_{11}) \middle| T_{11} = s \right\} h_1(s) \right. \\ &\quad \left. - E\left\{ \frac{n_1}{N} W_1(Z_{21}) + \frac{n_2}{N} W_2(Z_{21}) \middle| T_{21} = s - \theta(s) \right\} h_2\{s - \theta(s)\} \right] ds, \\ E(U_\phi) &= \int_{-\infty}^{\tau} \dot{\omega}(t) \left[E\left\{ \frac{n_1}{N} G_1(Z_{11}) + \frac{n_2}{N} G_2(Z_{11}) \middle| T_{11} = s \right\} h_1(s) \right. \\ &\quad \left. - E\left\{ \frac{n_1}{N} G_1(Z_{21}) + \frac{n_2}{N} G_2(Z_{21}) \middle| T_{21} = s - \theta(s) \right\} h_2\{s - \theta(s)\} \right] ds. \end{aligned}$$

The proof of Proposition 1 makes use of the following two lemmas.

Lemma 1

If C_k is independent of (T_k, Z_k) , then

$$P(X_k \leq t, \delta_k = 1, Z_k \leq z) = \int_{-\infty}^t S_k^O(r) \int_{-\infty}^z \bar{f}_k(r, u) du dr.$$

Lemma 2

If (Z_1, X_1, δ_1) and (Z_2, X_2, δ_2) are independent random triples and C_k is independent of (T_k, Z_k) for $k = 1, 2$, then

$$P(Z_2 \leq Z_1, X_1 \leq t, \delta_1 = 1, \delta_2 = 1) = \int_{-\infty}^t \left\{ S_1^O(r) \int_{-\infty}^{\infty} W_2(z) \bar{f}_1(r, z) dz \right\} dr.$$

Lemma 1 is adapted from Gilbert *et al.*, (2004, Appendix, Lemma1), and Lemma 2 is used for rank-based tests. The proofs of Lemmas 1–2 are straightforward but tedious and therefore omitted. Proof of Proposition 1 follows from these lemmas.

Next, we can establish the asymptotic distribution of L , U_1 and U_ϕ by exploiting empirical process theory; for example, Kosorok (2008, § 4.2.1). Now, we state the result for U_1 in Theorem 1.

Theorem 1

Assume that the proportion $n_1/N \rightarrow \rho$ as $N \rightarrow \infty$. Under conditions A1–A2, $N^{1/2}(U_1 - \vartheta_1)$ converges in distribution to a mean-zero normal random variable with finite variance, where

$$\begin{aligned} \vartheta_1 = & \int_{-\infty}^{\tau} \dot{\omega}(t) [E\{\rho W_1(Z_{11}) + (1 - \rho)W_2(Z_{11}) | T_{11} = s\} h_1(s) \\ & - E\{\rho W_1(Z_{21}) + (1 - \rho)W_2(Z_{21}) | T_{21} = s - \theta(s)\} h_2\{s - \theta(s)\}] ds. \end{aligned}$$

Under the null hypothesis in (2), $\vartheta_1 = 0$.

Proof of Theorem 1

Let $\mathcal{U}_1$ be an arbitrary compact interval containing ϑ_1 . Under Conditions A1–A2, functions of the following class of processes are Donsker: $\{I(X_k - t)\delta_k, t \in [\tau_b, \tau_u]\}$, $\{I(Z_k - z)\delta_k, z \in \kappa_b, \kappa_u\}$, $\{I(X_k - t, Z_k - z)\delta_k, (t, z) \in [\tau_b, \tau_u] \times [\kappa_b, \kappa_u]\}$, $\{I(X_k - t), t \in [\tau_b, \tau_u]\}$. Hence, the class of functions forms a Glivenko–Cantelli class on $[\tau_b, \tau_u] \times [\kappa_b, \kappa_u]$. By the Glivenko–Cantelli property, we have that each of $N_{Nk}(t) = n_k^{-1} \sum_{i=1}^{n_k} I(X_{ki} \leq t, \delta_{ki} = 1)$, $C_{Nk}(z)$, $B_{Nk}(t, z)$, $Y_{Nk}(t)$ converge in probability to their non-stochastic limits, uniformly in $[\tau_b, \tau_u] \times [\kappa_b, \kappa_u]$. Consequently, because the limit of $Y_{Nk}(t)$ is bounded in the succeeding texts by

Condition A1, U_1 converges in probability to $E(U_1)$, uniformly on $[\tau_b, \tau_u] \times [\kappa_b, \kappa_u]$. Again, by the Glivenko–Cantelli property, the empirical processes $\{N_{Nk}(\cdot), C_{Nk}(\cdot), B_{Nk}(\cdot, \cdot), Y_{Nk}(\cdot), k = 1, 2\}$ converge weakly to a Gaussian process $[\tau_b, \tau_u] \times [\kappa_b, \kappa_u]$. Now, the empirical average of marked ranked process is asymptotically equivalent to $Q_{Nk}(t) = \int \{\rho C_{N1}(z) + (1 - \rho)C_{N2}(z)\} dB_{Nk}(t, dz)$. By direct substitution, $t - \hat{\theta}_N(t) = S_{N2}^{-1}\{S_{N1}(t)\}$, where by Condition A1, $S_{N2}^{-1}(p)$, $p \in [a, b] \subset (0, 1)$ for some a, b and $S_{Nk}(\cdot) = \exp\{-H_{Nk}(\cdot)\}$ for $H_{Nk}(t) = \int_{-\infty}^t \{Y_{Nk}(u)\}^{-1} dN_{Nk}(u)$. One can then show that U_1 , with $\omega(t)$ in Condition A2 replacing $\omega(t)$, is a Hadamard-differentiable functional of $\{Q_{Nk}(\cdot), Y_{Nk}(\cdot), \hat{\theta}_N(\cdot), k = 1, 2\}$, and $Q_{Nk}(\cdot)$ and $\hat{\theta}_N(\cdot)$ are Hadamard-differentiable functionals of $\{N_{Nk}(\cdot), C_{Nk}(\cdot), B_{Nk}(\cdot, \cdot), Y_{Nk}(\cdot), k = 1, 2\}$. Thus, $(U_1 - \vartheta_1)$ is asymptotically normal by the functional delta method. Under H_0 , U_1 has mean zero by Proposition 1.

Using analogous arguments, one can show that the statistics L and U_ϕ are asymptotically normal, and the proposed tests are thus consistent.

References

- Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika*. 2000; 87:329–343.
- Fleming, TR, , Harrington, DP. Counting processes and survival analysis Wiley; New York: 1990
- Gelber RD, Gelman RS, Goldhirsch A. A quality-of-life oriented endpoint for comparing therapies. *Biometrics*. 1989; 45:781–795. [PubMed: 2790121]
- Genest C. Frank's family of bivariate distributions. *Biometrika*. 1987; 74:549–555.
- Gilbert PB, McKeague IW, Sun Y. Tests for comparing mark-specific hazards and cumulative incidence functions. *Lifetime Data Analysis*. 2004; 10:5–28. [PubMed: 15130048]
- Gilbert PB, McKeague IW, Sun Y. The 2-sample problem for failure rates depending on a continuous mark: An application to vaccine efficacy. *Biostatistics*. 2008; 9:263–276. [PubMed: 17704528]
- Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. *Statist Med*. 1990; 9:1259–76.
- Huang Y. Calibration regression of censored lifetime medical costs. *J Am Statist Assoc*. 2002; 97:318–327.
- Huang Y. Cost analysis with censored data. *Medical Care*. 2009; 47:S115–S119. [PubMed: 19536024]
- Huang Y, Berry K. Semiparametric estimation of the marginal mark distribution. *Biometrika*. 2006; 93:895–910.
- Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*. 1998; 85:785–798.
- Huang Y, Lovato L. Tests for lifetime utility or cost via calibrating survival time. *Stat Sinica*. 2002; 12:707–723.
- Kalbfleisch, JD, , Prentice, RL. The statistical analysis of failure time data John Wiley & Sons, Inc; Hoboken New Jersey: 2002
- Kosorok, MR. Introduction to empirical processes and semiparametric inference Springer; New York: 2008
- Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up. *Biometrics*. 1997; 53:419–434. [PubMed: 9192444]
- Strawderman RL. Estimating the mean of an increasing stochastic process at a censored stopping time. *J Am Statist Assoc*. 2002; 95:1192–1208.
- Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. *Ann Statist*. 1990; 18:354–372.
- Tsiatis, AA. Semiparametric theory and missing data Springer; New York: 2006

- Wang, JL. Smoothing hazard rates. In: Armitage, P, , Colton, T, editors Encyclopedia of Biostatistics
Wiley; New York: 2005
- Zhao H, Tsiatis AA. A consistent estimator for the distribution of quality adjusted survival time.
Biometrika. 1997; 84:339–348.

Fig. 1.

Empirical size of two-sample tests at the nominal 5% level where T_k and Z_k have marginal distributions F_k and G_k , respectively, $F_2(t) = F_1(t - 0.2)$, $F_1(t)$ is standard extreme value distribution, $G_2(z) = G_1(z)$ is standard exponential, (T_k, Z_k) has joint distribution $\bar{F}_k, \bar{F}_1(t, z)$ follows Frank's bivariate copula with association given by Kendall's τ . Each data point in every panel of the figure represents empirical size computed from 5000 Monte Carlo data sets. [Colour figure can be viewed at wileyonlinelibrary.com]

Fig. 2.

Empirical power of two-sample tests at the nominal 5% level where T_k and Z_k have marginal distributions F_k and G_k , respectively, $F_2(t) = F_1(t - 0.2)$, $F_1(t)$ is standard extreme value distribution, $G_2(z) = G_1(z - \eta)$, $G_1(z)$ is standard uniform with $\eta = 0.1$, (T_k, Z_k) has joint distribution $\bar{F}_k$, $\bar{F}_1(t, z)$ follows Frank's bivariate copula with association given by Kendall's τ . Each data point in every panel of the figure represents empirical power computed from 5000 Monte Carlo data sets. [Colour figure can be viewed at wileyonlinelibrary.com]

Fig. 3.

Empirical power of two-sample tests at the nominal 5% level where T_k and Z_k have marginal distributions F_k and G_k , respectively, $F_2(t) = F_1(t - 0.2)$, $F_1(t)$ is standard extreme value distribution, $G_2(z) = G_1(z - \eta)$, $G_1(z)$ is standard exponential with $\eta = 0.25$, (T_k, Z_k) has joint distribution $\bar{F}_k$, $\bar{F}_1(t, z)$ follows Frank's bivariate copula with association given by

Kendall's τ . Each data point in every panel of the figure represents empirical power computed from 5000 Monte Carlo data sets. [Colour figure can be viewed at

wileyonlinelibrary.com]

Fig. 4.

Analysis results from Southwest Oncology Group Study 9509. In each panel, the heavy grey line is the estimated stochastic process for the Paclitaxel plus Carboplatin group while the black line is an estimate for the Vinorelbine plus Cisplatin group. The Kaplan–Meier estimates of the survival probability are displayed in panel (a); the cumulative marked hazard functions are displayed in panels (b)–(d) for the marked identity $A_{NK}(t)$, rank $Q_{NK}(t)$ and weighted rank $Q_{NK}(t, \phi)$ processes, respectively.

Fig. 5.

Histograms of 1000 perturbed test statistics U_1^* (Rank), U_ϕ^* (Wt.Rank) and L^* (Identity) where lifetime medical cost is on the original scale.

Details of test statistics for comparing cumulative marked hazard rates of lifetime medical cost in Southwest Oncology Group Study 9509

Table 1

Method	$\omega(p)$	Numerator	SD	Rob.SD	z-score	p-value
<u>Medical cost on a logarithmic scale</u>						
Identity (L)	Gehan	-36.18	12.48	12.92	-2.90	0.0039
	log-rank	-50.87	18.49	18.92	-2.75	0.0079
Rank (U_l)	Gehan	-7.90	2.83	2.95	-2.79	0.0054
	log-rank	-12.59	4.65	4.55	-2.71	0.0078
Wt.Rank (U_ϕ)	Gehan	-8.77	3.20	3.32	-2.74	0.0062
	log-rank	-14.13	5.34	5.26	-2.65	0.0092
<u>Medical cost on the original scale</u>						
Identity (L)	Gehan	-0.73	0.31	0.32	-2.36	0.0186
	log-rank	-1.25	0.57	0.54	-2.18	0.0312

Table entries include the numerators (U_l , U_ϕ or L), estimated standard deviation (SD) and robust standard deviation (Rob.SD) divided by 1000. We also report the test statistic as the z-score equal to the numerator divided by SD and the p-value for a two-sided test of the null hypothesis.