Ann Inst Stat Math (2013) 65:149–166
DOI 10.1007/s10463-012-0378-5

On Mann–Whitney tests for comparing sojourn time distributions when the transition times are right censored

Jie Fan · Somnath Datta

Received: 10 February 2011 / Revised: 20 February 2012 / Published online: 7 October 2012

Abstract We consider the problem of comparing sojourn time distributions of a transient state in a general multistate system in two samples (groups) when the transition times are right censored. Using the reweighting principle, a two-sample Mann–Whitney type of U-statistic is constructed that compares only the uncensored sojourn times from the two distributions. A second Mann–Whitney type of statistic is also constructed using a different reweighting that allows for comparisons when one of the two sojourn times is either uncensored or singly censored. Both these statistics are asymptotically unbiased, asymptotically normally distributed and reduce to the standard Mann–Whitney statistic when there is no censoring. A test of equality of sojourn time distributions in two independent samples is constructed by symmetrizing the second statistic. The testing methodology is illustrated using a data set on kidney disease patients.

Keywords Censoring · Martingale · Mann–Whitney statistic · Reweighting principle · U-statistic · Waiting time

1 Introduction

The Mann–Whitney U test (Mann and Whitney 1947) is perhaps the most commonly used nonparametric procedure in comparing two distributions based on independent samples. Procedurally, it is equivalent to Wilcoxon's rank-sum test and because of that test is sometimes collectively referred to as Wilcoxon–Mann–Whitney test. The test was proposed initially by Wilcoxon (1945) for equal sample sizes in two groups and later extended by Mann and Whitney (1947) for possibly unequal sample sizes.

J. Fan · S. Datta (✉)
Department of Biostatistics and Bioinformatics, University of Louisville, Louisville, KY 40202, USA
e-mail: Somnath.Datta@louisville.edu

Practitioners often regard the Mann–Whitney test as the nonparametric counterpart of the parametric two-sample t test. However, the method is more robust than the t test and can be applied to ordinal data in addition to continuous data. It is especially useful when the assumption of normality is not met.

Unfortunately, the traditional Mann–Whitney U test does not take missing values into account. In this paper, we are interested in comparing the sojourn times (waiting times) in two independent samples when the transition times (e.g., both the state entry and the state exit times) are subject to right censoring. In this situation, missing data could arise if at least one of the state entry or exit times is right censored. Unlike right-censored failure time data (Latta 1977; Prentice 1978, etc.), there are currently no extensions of the Wilcoxon–Mann–Whitney test that applies to this situation since the censoring induced on the set of sojourn times is more complex than independent right censoring.

Our attempts to extend the Mann–Whitney U-statistic to sojourn times under right-censored transition times are based on its representation as a generalized U-statistic. Traditional U-statistics (Serfling 1980) are one-sample statistics, whereas generalized U-statistics are based on k (≥ 2) samples $\{X_{1,1}, \ldots X_{n_1,1}\}, \ldots, \{X_{1,k}, \ldots X_{n_k,k}\}$ from distributions $F_1, \ldots, F_k$, respectively. See, e.g., Serfling (1980, p. 175) for a formal definition.

Datta et al. (2010) proposed an inverse probability of censoring weighted (IPCW) U-statistics for right-censored data. Earlier, Schisterman and Rotnizky (2001) considered inverse probability weighting for constructing U-statistics based on missing data. We adopt similar reweighting principles to deal with missing sojourn times. However, the current setup is more complicated for two reasons. First of all, unlike U-statistics, generalized U-statistics involve multiple groups and, more importantly, the right-censoring mechanism operates on the transition times and not on the sojourn times inducing a dependent censoring.

The rest of the paper is organized as follows. In Sect. 2, we introduce the two proposed Mann–Whitney type of U-statistics to compare two sojourn time distributions when transition times are right censored. In the third section, we describe the asymptotic properties of the proposed statistics including variance estimation. In Sect. 4, we consider testing the null hypothesis of equality of sojourn time distributions from two independent samples. We construct our test statistic using a symmetrization of a Mann–Whitney type of U-statistic introduced by us. We present results from a number of simulation studies for both estimation and testing by generating data from different scenarios in Sect. 5. In Sect. 6, we apply the testing methodology to a kidney disease data set as an illustration. We conclude the main body of the paper with discussion in Sect. 7. The proofs of asymptotic normality of the two Mann–Whitney type of statistics are placed in the Appendix.

2 Mann–Whitney U-statistics for sojourn times in the presence of right censoring

We begin this section by introducing the notation necessary to describe our statistics. Suppose we have right-censored entry and exit time data from two independent populations (groups). Let $X^*_{i,j}$ and $V^*_{i,j}$ be the possibly unobserved (due to right

censoring) state entry and exit times, respectively, for the ith subject in the jth group, both of which are subject to right censoring by a common censoring time $C_{i,j}$ which is assumed to be independent of the pair $(X^*_{i,j}, V^*_{i,j})$. Our observed data consist of the four tuples $(X_{i,j}, \xi_{i,j}, V_{i,j}, \delta_{i,j})$, $1 \le i \le n_j$, $j = 1, 2$, where $X_{i,j} = \min(X^*_{i,j}, C_{i,j})$ and $V_{i,j} = \min(V^*_{i,j}, C_{i,j})$ are the (right) censored state entry and exit times, and $\xi_{i,j} = I(X^*_{i,j} \le C_{i,j})$ and $\delta_{i,j} = I(V^*_{i,j} \le C_{i,j})$ are the censoring indicators for the ith subject in the jth group. Also let, for future use, $\overline{\xi_{i,j}} = 1 - \xi_{i,j}$ and $\overline{\delta_{i,j}} = 1 - \delta_{i,j}$. Let $W^*_{i,j} = V^*_{i,j} - X^*_{i,j}$ be the possibly unobserved sojourn times and we define $W_{i,j} = V_{i,j} - X_{i,j}$. Note that $W_{i,j}$ is computable from the observed data and equals $W^*_{i,j}$ if and only if $\delta_{i,j} = 1$. Let F_j be the sojourn time distribution function in group j and $S_j = 1 - F_j$.

When we compare two sojourn time distributions with no missing observations, the Mann–Whitney U-statistic is given by $U^* = (n_1 n_2)^{-1} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} I(W^*_{i_1,1} \le W^*_{i_2,2})$. In the present context, we replace W^* by the observed data quantities W for each pair with both $\delta = 1$; i.e., we only select the fully observed sojourn times from each group for comparison. In order to compensate for this selection bias, we reweigh each summand by the inverse of the selection probabilities conditional on the state exit times for such a sample pair leading to the following extension of Mann–Whitney statistic

$$U_1 = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{I(W_{i_1,1} \le W_{i_2,2}) \delta_{i_1,1} \delta_{i_2,2}}{K_1(V_{i_1,1}-) K_2(V_{i_2,2}-)},$$

where $K_j(t) = P\{C_j > t\}$ is the survival function of the censoring times in group j.

The following simple argument shows that indeed U_1 agrees with the full data Mann–Whitney statistic U^* on average:

$$\begin{aligned} E(U_1) &= \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} E\left[E\left\{ \frac{I(W^*_{i_1,1} \le W^*_{i_2,2}) \delta_{i_1,1} \delta_{i_2,2}}{K_1(V^*_{i_1,1}-) K_2(V^*_{i_2,2}-)} \middle| X^*_{i_1,1}, V^*_{i_1,1}, X^*_{i_2,2}, V^*_{i_2,2} \right\} \right] \\ &= \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} E\left\{ \frac{I(W^*_{i_1,1} \le W^*_{i_2,2})}{K_1(V^*_{i_1,1}-) K_2(V^*_{i_2,2}-)} P(C_{i_1,1} \ge V^*_{i_1,1} | X^*_{i_1,1}, V^*_{i_1,1}) \right. \\ &\qquad \left. \times P(C_{i_2,1} \ge V^*_{i_2,1} | X^*_{i_2,2}, V^*_{i_2,2}) \right\} \end{aligned}$$

by independence of two samples,

$$= \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} E\left\{ \frac{I(W^*_{i_1,1} \le W^*_{i_2,2})}{K_1(V^*_{i_1,1}-) K_2(V^*_{i_2,2}-)} K_1(V^*_{i_1,1}-) K_2(V^*_{i_2,2}-) \right\},$$

by independence of C_{ij} and $\{X^*_{ij}, V^*_{ij}\}$,

$$= \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} E\left\{ I(W^*_{i_1,1} \le W^*_{i_2,2}) \right\} = E(U^*).$$

Note that U_1 is not a statistic in the strict sense of the word since it involves the population quantities K_1 and K_2. We estimate K_j by the group-specific Kaplan–Meier estimator $\widehat{K}_j$ for the censoring survival function. Note that $\widehat{K}_j(t-)$ can be computed based on sample $j(= 1, 2)$ by the standard Kaplan–Meier formula where the roles of failure and censoring times are switched and a C_{ij} exceeding the corresponding exit times $V^*_{i,j}$ is considered to be censored. Substituting $\widehat{K}_j$ in place of K_j, we get our first Mann–Whitney type statistic

$$\widehat{U}_1 = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{I(W_{i_1,1} \le W_{i_2,2}) \delta_{i_1,1} \delta_{i_2,2}}{\widehat{K}_1(V_{i_1,1}-) \widehat{K}_2(V_{i_2,2}-)}.$$

Next, we propose a second generalization of Mann–Whitney U-statistic for sojourn times that allows for comparison of additional pairs even when they are not fully observed. Note that the indicator kernel $I(W^*_{i_1,1} \le W^*_{i_2,2})$ can be evaluated when $W^*_{i_1,1} = W_{i_1,1}$ is non-missing and we can conclude that $W^*_{i_2,2}$ is larger than $W_{i_1,1}$ from the fact that $W_{i_2,2}$ is larger than $W_{i_1,1}$. In other words, the second entry time $X^*_{i_2,2} = X_{i_2,2}$ has to be non-missing as well and the second censoring time is at least $W_{i_1,1} + X_{i_2,2}$. The probability of both of these events occurring together given $\{X^*_{i_1,1}, V^*_{i_1,1}, X^*_{i_2,2}, V^*_{i_2,2}\}$ is $K_1(V^*_{i_1,1}-)K_2(W^*_{i_1,1} + X^*_{i_2,2}-)$, which is the same as $K_1(V_{i_1,1}-)K_2(W_{i_1,1} + X_{i_2,2}-)$ on the set $\delta_{i_1,1}\xi_{i_2,2} = 1$. Thus, we obtain our second generalization of Mann–Whitney sojourn times statistic

$$\widehat{U}_2 = \frac{1}{n_1 n_2} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{I(W_{i_1,1} \le W_{i_2,2}) \delta_{i_1,1} \xi_{i_2,2}}{\widehat{K}_1(V_{i_1,1}-) \widehat{K}_2(W_{i_1,1} + X_{i_2,2}-)}. \tag{1}$$

Its unbiasedness, with the true K_j in place of $\widehat{K}_j$, can be established as before. We expect it to be more efficient than $\widehat{U}_1$, since it is based on non-zero scores on a larger number of pairs.

3 Large sample properties

We denote the population quantity $P\{W^*_1 \le W^*_2\}$ estimated by the Mann–Whitney U-statistics by θ. Note that $\theta \in [0, 1]$ provides a measure of stochastic order between the two continuous sojourn time distributions. In particular, $\theta <, =, > 1/2$, when the sojourn time distribution in the first group is stochastically larger than, equal to, or smaller than the sojourn time distribution in the second group, respectively. Besides testing the equality of two distributions, the Mann–Whitney statistic is also useful for providing a point estimate of θ.

We need to introduce the following counting process notation (see., e.g., Andersen et al. 1993). Let $N^c_{i,j}(t) = I(V_{i,j} \le t, \delta_{i,j} = 0)$ be the counting processes of

censoring, $Y_{i,j}(t) = I(V_{i,j} \geq t)$ be the "number at-risk" processes, and $M^c_{i,j}(t) = N^c_{i,j}(t) - \int_0^t Y_{i,j}(u)\mathrm{d}\Lambda^c_j(u)$ be the martingale of the censoring process defined with respect to the appropriate filtration for the two samples; here, Λ^c_j is the cumulative hazard for censoring in the jth group, $j = 1, 2$. Let $Y_j(t) = \sum_{i=1}^n Y_{ij}(t),\ j = 1, 2$, and let $\overline{n}_j$ be the sub-distribution function of the pair (W_j, V_j) corresponding to $\delta_j = 1,\ j = 1,\ 2$, and $\overline{n}_3$ be the sub-distribution function of the pair (W_2, X_2) corresponding to $\xi_2 = 1$,

$$\overline{n}_j(w, v) = P\{W_j \leq w, V_j \leq v,\ \delta_j = 1\},\ j = 1, 2,$$
$$\overline{n}_3(w, x) = P\{W_2 \leq w, X_2 \leq x,\ \xi_2 = 1\}.$$

Consider the following univariate functions on $[0, \infty)$:

$$\omega_1(s) = \frac{1}{y_1(s)} \int I(v > s) \frac{S_2(w)}{K_1(v-)} \mathrm{d}\overline{n}_1(w, v), \tag{2}$$

$$\omega_2(s) = \frac{1}{y_2(s)} \int I(v > s) \frac{F_1(w-)}{K_2(v-)} \mathrm{d}\overline{n}_2(w, v), \tag{3}$$

$$\omega_3(s) = \frac{1}{y_2(s)} \int \frac{I(w_1 + x_2 > s) I(w_1 < w_2) \mathrm{d}F_1(w_1) \mathrm{d}\overline{n}_3(w_2, x_2)}{K_2(w_1 + x_2-)} \tag{4}$$

where $y_j(s) = P(V_j \geq s),\ j = 1, 2,\ s \geq 0$.

Theorem 1 *Under suitable regularity conditions (see the Appendix), as* $n \to \infty$, *where* $n = n_1 + n_2$, *we get* $\sqrt{n}(\widehat{U}_1 - \theta) \xrightarrow{d} N(0, \sigma_1^2)$, *where* $\sigma_1^2 = c_1^{-1} var\{S_2(W_1)\delta_1/K_1(V_1-) + \int_0^\infty \omega_1(s)\mathrm{d}M^c_1(s)\} + c_2^{-1} var\{F_1(W_2-)\delta_2/K_2(V_2-) + \int_0^\infty \omega_2(s)\mathrm{d}M^c_2(s)\}$, *and* $\sqrt{n}(\widehat{U}_2 - \theta) \xrightarrow{d} N(0, \sigma_2^2)$, *where* $\sigma_2^2 = c_1^{-1} var\{S_2(W_1)\delta_1/K_1(V_1-) + \int_0^\infty \omega_1(s)\mathrm{d}M^c_1(s)\} + c_2^{-1} var[\xi_2 \int_0^\infty \{I(w \leq W_2)/K_2(w + X_2-)\}\mathrm{d}F_1(w) + \int_0^\infty \omega_3(s)\mathrm{d}M^c_2(s)]$, *with* $c_j = \lim(n_j/n),\ j = 1, 2$.

The above expressions for the asymptotic variances also suggest the following natural estimators:

$$\widehat{\sigma}_1^2 = \frac{n}{n_1(n_1 - 1)} \sum_{i_1=1}^{n_1} (S_{i_1,1} - \overline{S}_1)^2 + \frac{n}{n_2(n_2 - 1)} \sum_{i_2=1}^{n_2} (S_{i_2,2} - \overline{S}_2)^2,$$

where

$$S_{i,1} = \frac{\widehat{S}_2(W_{i,1})\delta_{i,1}}{\widehat{K}_1(V_{i,1}-)} + \widehat{\omega}_1(V_{i,1})\overline{\delta}_{i,1} - \sum_{i_1=1}^{n_1} \frac{\widehat{\omega}_1(V_{i_1,1})\, I(V_{i,1} \geq V_{i_1,1})\overline{\delta}_{i_i,1}}{Y_1(V_{i_1,1})}, \tag{5}$$

$$S_{i,2} = \frac{\widehat{F}_1(W_{i,2})\delta_{i,2}}{\widehat{K}_2(V_{i,2}-)} + \widehat{\omega}_2(V_{i,2})\overline{\delta}_{i,2} - \sum_{i_2=1}^{n_2} \frac{\widehat{\omega}_2(V_{i_2,2})\, I(V_{i,2} \geq V_{i_2,2})\overline{\delta}_{i_2,2}}{Y_2(V_{i_2,2})}, \tag{6}$$

$\overline{S}_j = n_j^{-1} \sum_{i=1}^{n_j} S_{ij}$, $j > 1$, and where

$$\widehat{\omega}_1(s) = \frac{1}{Y_1(s)} \sum_{i=1}^{n_1} I(V_{i,1} > s) \frac{\widehat{S}_2(W_{i,1})\delta_{i,1}}{\widehat{K}_1(V_{i,1}-)}, \tag{7}$$

$$\widehat{\omega}_2(s) = \frac{1}{Y_2(s)} \sum_{i=1}^{n_2} I(V_{i,2} > s) \frac{\widehat{F}_1(W_{i,2}-)\delta_{i,2}}{\widehat{K}_2(V_{i,2}-)}. \tag{8}$$

In (5) and (7) above, $\widehat{S}_2$ is the Satten–Datta estimator (Satten and Datta 2002) of the survival function of W_2^* based on sample 2, i.e.,

$$\widehat{S}_2(W_{i,1}) = \frac{1}{n_2} \sum_{i_2=1}^{n_2} \frac{I(W_{i_2,2} > W_{i,1})\delta_{i_2,2}}{\widehat{K}_2(V_{i_2,2}-)};$$

in (6) and (8), $\widehat{F}_1 = 1 - \widehat{S}_1$ is the Satten–Datta estimator of the distribution function of W_1^* based on sample 1, i.e.,

$$\widehat{F}_1(W_{i,2}) = \frac{1}{n_1} \sum_{i_1=1}^{n_1} \frac{I(W_{i_1,1} \le W_{i,2})\delta_{i_1,1}}{\widehat{K}_1(V_{i_1,1}-)}.$$

Similarly, σ_2^2 can be estimated by

$$\widehat{\sigma}_2^2 = \frac{n}{n_1(n_1-1)} \sum_{i_1=1}^{n_1} (S_{i_1,1} - \overline{S}_1)^2 + \frac{n}{n_2(n_2-1)} \sum_{i_2=1}^{n_2} (S_{i_2,3} - \overline{S}_3)^2$$

with

$$S_{i,3} = \xi_{i,2} \frac{1}{n_1} \sum_{i_1=1}^{n_1} \frac{I(W_{i_1,1} \le W_{i,2})\delta_{i_1,1}}{\widehat{K}_1(V_{i_1,1}-)\widehat{K}_2(W_{i_1,1} + X_{i,2}-)} + \widehat{\omega}_3(V_{i,2})\overline{\xi}_{i,2} - \sum_{i_2=1}^{n_2} \frac{\widehat{\omega}_3(V_{i_2,2})\, I(V_{i,2} \ge V_{i_2,2})\overline{\xi}_{i_2,2}}{Y_2(V_{i_2,2})},$$

and

$$\widehat{\omega}_3(s) = \frac{1}{Y_2(s)} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{\delta_{i_1,1}\xi_{i_2,2} I(W_{i_1,1} + X_{i_2,2} > s) I(W_{i_1,1} < W_{i_2,2})}{\widehat{K}_1(V_{i_1,1}-)\widehat{K}_2(W_{i_1,1} + X_{i_2,2}-)}$$

with $\overline{\xi}_{i_2,2} = 1 - \xi_{i_2,2}$.

Consistency of the above variance estimators can be shown using projection techniques for generalized U-statistics and results for reweighting as in the proof of Theorem 1.

Remark 1 Even though the computation of $\widehat{U}_2$ is more involved than $\widehat{U}_1$, we expect it to be more efficient since it effectively uses a larger number of sample pairs. In simulation studies reported in Sect. 5, we note that indeed $\widehat{U}_2$ has a slightly smaller variance than $\widehat{U}_1$, in all the settings that were tried while the biases were comparable. Therefore, we only consider $\widehat{U}_2$ for constructing our test statistic in the following section.

4 Testing the equality of sojourn time distributions in 2 groups

We now use the second Mann–Whitney type of statistic $\widehat{U}_2$ for testing the equality of sojourn time distributions in two groups based on independent samples from these groups in the presence of right censoring on the transition times. We assume that the sojourn time distributions are continuous. In this section, we denote by $\widehat{U}(1,2)$ the test-statistic $\widehat{U}_2$ given in (1) based on group 1 and group 2 samples in that order. Note that, for our censored setup, the statistics $\widehat{U}(1,2)$ and $1-\widehat{U}(2,1)$ will be close, but not equal unless all sojourn times are non-missing for a sample. Therefore, we could take their average $T = 0.5\{\widehat{U}(1,2) + 1 - \widehat{U}(2,1)\}$ as the (one-sided) test statistic for testing the null hypothesis $H_0 : F_1 = F_2$ of equality of group 1 and 2 sojourn time distributions. Under the null hypothesis, T has an asymptotic mean of $\theta = 0.5$. Following the same linearizations as in Theorem 1, we can estimate its asymptotic variance $n^{-1}\widehat{\sigma}^2_{H_0}$ by

$$\widehat{\sigma}^2_{H_0} = \frac{n}{4n_1(n_1-1)}\sum_{i_1=1}^{n_1}(S_{i_1,4}-\overline{S}_4)^2 + \frac{n}{4n_2(n_2-1)}\sum_{i_2=1}^{n_2}(S_{i_2,5}-\overline{S}_5)^2,$$

where

$$\begin{aligned} S_{i,4} = {} & \frac{\widehat{S}_2(W_{i,1})\delta_{i,1}}{\widehat{K}_1(V_{i,1}-)} + \widehat{\omega}_1(V_{i,1})\overline{\delta}_{i,1} - \sum_{i_1=1}^{n_1}\frac{\widehat{\omega}_1(V_{i_1,1})\, I(V_{i,1}\geq V_{i_1,1})\overline{\delta}_{i,1}}{Y_1(V_{i_1,1})} \\ & - \frac{\xi_{i,1}}{n_2}\sum_{i_2=1}^{n_2}\frac{I(W_{i_2,1}\leq W_{i,2})\delta_{i_2,1}}{\widehat{K}_2(V_{i_2,1}-)\widehat{K}_1(W_{i_2,1}+X_{i,2}-)} - \widehat{\omega}_4(V_{i,1})\overline{\xi}_{i,1} \\ & + \sum_{i_1=1}^{n_1}\frac{\widehat{\omega}_4(V_{i_1,1})\, I(V_{i,1}\geq V_{i_1,1})\overline{\xi}_{i_1,1}}{Y_1(V_{i_1,1})} \end{aligned}$$

and

$$S_{i,5} = \frac{\xi_{i,2}}{n_1}\sum_{i_1=1}^{n_1}\frac{I(W_{i_1,1}\leq W_{i,2})\delta_{i_1,1}}{\widehat{K}_1(V_{i_1,1}-)\widehat{K}_2(W_{i_1,1}+X_{i,2}-)} + \widehat{\omega}_3(V_{i,2})\overline{\xi}_{i,2}$$

$$-\sum_{i_2=1}^{n_2} \frac{\widehat{\omega}_3(V_{i_2,2})\, I(V_{i,2} \geq V_{i_2,2})\overline{\xi}_{i_2,2}}{Y_2(V_{i_2,2})}$$

$$-\frac{\widehat{S}_1(W_{i,2})\delta_{i,2}}{\widehat{K}_2(V_{i,2}-)} - \widehat{\omega}_5(V_{i,2})\overline{\delta}_{i,2} + \sum_{i_2=1}^{n_2} \frac{\widehat{\omega}_5(V_{i_1,2})\, I(V_{i,2} \geq V_{i_2,2})\overline{\delta}_{i_2,2}}{Y_2(V_{i_2,2})}$$

with

$$\widehat{\omega}_4(s) = \frac{1}{Y_1(s)} \sum_{i_1=1}^{n_1} \sum_{i_2=1}^{n_2} \frac{\xi_{i_1,1}\delta_{i_2,2} I(X_{i_1,1} + W_{i_2,2} > s) I(W_{i_2,2} < W_{i_1,1})}{\widehat{K}_1(X_{i_1,1} + W_{i_2,2}-)\widehat{K}_2(V_{i_2,2}-)}$$

and

$$\widehat{\omega}_5(s) = \frac{1}{Y_2(s)} \sum_{i=1}^{n_2} I(V_{i,2} > s) \frac{\widehat{S}_1(W_{i,2})\delta_{i,2}}{\widehat{K}_2(V_{i,2}-)}.$$

Theorem 2 *Under the null hypothesis* $H_0 : F_1 = F_2$, $Z := \sqrt{n}\widehat{\sigma}_{H_0}^{-1}(T - 0.5) \xrightarrow{d} N(0, 1)$, *as* $n \to \infty$, *provided the regularity conditions of Theorem 1 hold.*

An empirical power study of this test is carried out in the second part of Sect. 5 to investigate the performance of this test in small to moderate samples.

Remark 2 One could apply the classical Mann–Whitney test using the censored W_{ij} disregarding the fact that some of them are censored. Since their group-specific distribution is a functional of the distribution of true W^* and the censoring variable, this naive approach may indeed be valid provided the censoring distributions (patterns) are the same in the two groups. However, this could lead to a substantial loss in power. Furthermore, when the censoring distributions in the two groups differ, this test may inflate the size. We demonstrate these very clearly in a simulation study reported in the next section.

5 Simulation studies

We conducted a number of simulation studies for investigating the finite sample behaviors of the Mann–Whitney type of statistics (Sect. 3) and the large sample test (Sect. 4) for the equality of two sojourn time distributions.

5.1 A semi-Markov model

In this simulation scenario, we generated sojourn times independently of the state entry times. The same distributions were used in both groups leading to $\theta = 0.5$. The state entry and the sojourn times were each generated from a standard lognormal distribution. The censoring times are also generated from a lognormal distribution with unit scale parameter, but with possibly different log mean parameters in the two

Table 1 Simulation results for U-statistics in a semi-Markov model when the censoring rates are different in two groups

Group size	Censoring rate 1/censoring rate 2						
	$\theta = 0.5$						$\theta = 0.9$
	0.25/0.25	0.50/0.50	0.75/0.75	0.25/0.5	0.25/0.75	0.5/0.75	0.5/0.5
25							
Bias($\widehat{U}_1$)	−0.011	−0.038	−0.096	−0.034	−0.082	−0.087	−0.174
Bias($\widehat{U}_2$)	−0.016	−0.042	−0.105	−0.038	−0.092	−0.097	−0.092
ESE($\widehat{U}_1$)	0.099	0.132	0.197	0.119	0.175	0.181	0.141
ESE($\widehat{U}_2$)	0.096	0.122	0.157	0.109	0.123	0.132	0.108
SE($\widehat{U}_1$)	0.098	0.125	0.169	0.115	0.143	0.151	0.169
SE($\widehat{U}_2$)	0.095	0.116	0.148	0.105	0.117	0.127	0.130
50							
Bias($\widehat{U}_1$)	−0.005	−0.019	−0.060	−0.017	−0.052	−0.053	−0.121
Bias($\widehat{U}_2$)	−0.008	−0.024	−0.060	−0.023	−0.052	−0.053	−0.064
ESE($\widehat{U}_1$)	0.070	0.094	0.139	0.086	0.122	0.128	0.106
ESE($\widehat{U}_2$)	0.068	0.086	0.119	0.077	0.096	0.104	0.080
SE($\widehat{U}_1$)	0.069	0.089	0.126	0.082	0.107	0.113	0.120
SE($\widehat{U}_2$)	0.067	0.083	0.111	0.075	0.088	0.095	0.091

SE standard error, *ESE* estimated standard error

groups which were varied to achieve different censoring rates. Here and subsequently, by censoring rates we refer to the probability $P(V_{ij}^* > C_{ij})$ of the exit times being right censored.

In all cases, equal sample sizes ($n_j = 25$ and 50) in two groups were used. The left columns (3–6) in Table 1 report the results when the same censoring rates were used in two groups. The common censoring rates varied from low (25 %) to heavy (75 %). A Monte Carlo size of 1,000 was used to compute the answers reported in Table 1. From Table 1, it is evident that the variance formulas work since the estimated standard errors are close to the empirical standard errors for both methods. Biases and standard errors increase for both methods when the censoring rate increases and/or the group sample size decreases, as expected. We also find that the bias for $\widehat{U}_1$ is very slightly smaller than $\widehat{U}_2$ under this simulation scenario for the smaller sample size; however, the estimated standard error for $\widehat{U}_1$ is consistently larger than that for $\widehat{U}_2$.

Results for different degrees of censoring in the two groups are reported in columns 7–9 of Table 1. They have similar patterns for the bias and the standard error for the two methods as in the cases of equal censoring rates.

For the sake of completeness, we also report some results for the case when the sojourn times in the two groups do not have the same distribution. More precisely, we generate the sojourn times for the second group from a lognormal distribution with a log mean value of 1.8, whereas the sojourn time distribution for group 1 is unchanged from the earlier setting. This yields a value of $\theta = 0.9$. For the sake of brevity, we

only report the results when the censoring rates are 50 % in both groups. The results are reported in the rightmost column of Table 1. Both estimators now exhibit greater biases and standard errors than before. Presumably, this is due to the fact that the calculation involves summands and estimated weights that are more toward the tail of a distribution. Interestingly, now $\widehat{U}_2$ beats $\widehat{U}_1$ both in terms of bias and variance.

Based on all these simulations, we can conclude that the second statistics $\widehat{U}_2$ is a better choice for extending the Mann–Whitney statistic to the current setup involving right censoring on the transition times.

5.2 A Markov model

In this simulation setting, we generate entry times within each group from a standard lognormal distribution. After obtaining an entry time X^*, for example, the corresponding exit time was obtained by the formula

$$V^* = D^{-1}[D(X^*) + U\{1 - D(X^*)\}],$$

where $D(\cdot)$ is the distribution function of the standard lognormal distribution, U is a number randomly generated from a uniform distribution in the interval $[0, 1]$ and D^{-1} denotes the quantile function of the standard lognormal distribution. Note that this ensures that $V^* \geq X^*$; furthermore, the resulting system is Markov and the transition hazard for V^* is also that of a standard lognormal. The censoring times were generated by the same mechanism as in simulation 1 where we varied the common log mean parameter to control the censoring rates.

Table 2 summarizes the results of this simulation. Once again, the estimated standard errors are close to their population counterparts; both bias and standard error decrease with the sample size and increase with censoring percentage. There is no consistent comparative patterns for the biases, but the estimated standard error for $\widehat{U}_2$ is still consistently smaller than that for $\widehat{U}_1$.

5.3 Testing hypotheses for equality of sojourn times

We also conducted simulations for a power study using the test Z described in Sect. 4. We also include the standard Mann–Whitney test based on the censored version of the sojourn times W_{ij} (see Remark 1 of the previous section) for comparison. A nominal size of 5 % was selected for all tests.

First, we consider a situation when the two groups have the same censoring distribution in the two groups. We varied log mean in $[-1.5, 1.5]$ with a step of 0.1 in group 2, while keeping the other parameters the same as those in the simulation 1 with the censoring rate equal to 25 % (under H_0); in particular, the group 1 sojourn times were generated from a standard lognormal distribution. To reduce computational burden, we computed the power at fewer values when the common group size was 50. We simulated 1,000 data sets under each parameter setting. For each generated sample data set, the studentized test statistic Z was applied. The empirical power of the test at each alternative parameter setting was calculated by the proportion of

Table 2 Simulation results for U-statistics under a Markov model; equal censoring rates

Group size	Censoring rate		
	0.25	0.5	0.75
25			
Bias($\widehat{U}_1$)	−0.015	−0.047	−0.123
Bias($\widehat{U}_2$)	−0.019	−0.049	−0.113
ESE($\widehat{U}_1$)	0.097	0.136	0.200
ESE($\widehat{U}_2$)	0.094	0.116	0.158
SE($\widehat{U}_1$)	0.098	0.125	0.168
SE($\widehat{U}_2$)	0.094	0.114	0.146
50			
Bias($\widehat{U}_1$)	−0.009	−0.029	−0.082
Bias($\widehat{U}_2$)	−0.010	−0.029	−0.072
ESE($\widehat{U}_1$)	0.066	0.090	0.146
ESE($\widehat{U}_2$)	0.065	0.082	0.117
SE($\widehat{U}_1$)	0.068	0.089	0.127
SE($\widehat{U}_2$)	0.067	0.082	0.111

Here, the true θ is 0.5
SE standard error, *ESE* estimated standard error

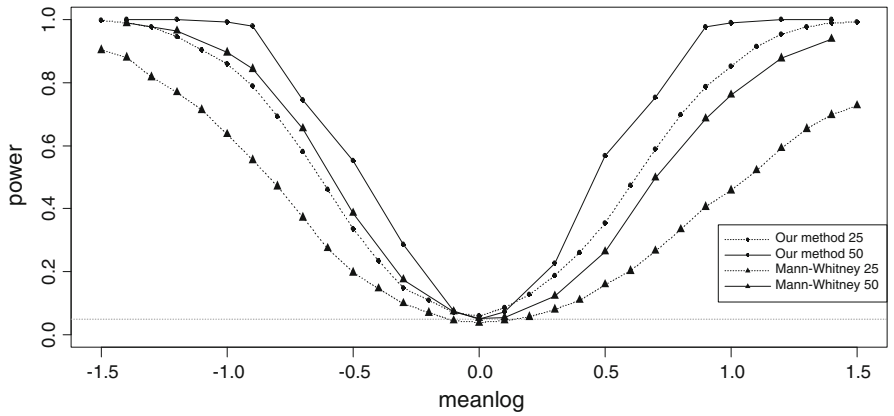

Fig. 1 Power plots for the proposed generalized Mann–Whitney test along with those for a naive Mann–Whitney tests comparing the censored sojourn times. The sojourn time distributions in both groups are lognormal with unit scale; the log mean in group 1 is 0; the power curves are plotted with respect to the log mean parameters in group 2. The censoring rate was kept at 25 % under the null hypothesis. The *dotted curve* corresponds to a sample size of 25 and the *solid curve* corresponds to sample size of 50 in each group, respectively

times the null hypothesis was rejected by this test out of 1,000 samples. Figure 1 displays the resulting power curves. While both tests maintained the nominal size, the power of our test is uniformly larger than the naive Mann–Whitney test for a given sample size. Power curves behave reasonably for both tests; in particular, they

are piecewise monotonic on $(-\infty, 0)$ and $(0, \infty)$ and the power increases with the sample size.

Next, we consider the setting as in Table 1, where the censoring rate in one group is 25 % and in the other group 50 % and, more importantly, the censoring distributions in the two groups were different. The sojourn times in the two groups have the same distributions. The size of the naive Mann–Whitney test was greatly inflated (the empirically estimated size was equal to 0.711). On the other hand, the size of our test was maintained at the nominal level. The empirically estimated size was 0.060 with a 95 % confidence interval of (0.045, 0.075). This example clearly demonstrates the danger of using a naive Mann–Whitney test and the utility of our modified Mann–Whitney test in comparing two sets of sojourn times.

6 An illustration using kidney disease data

McGilchrist and Aisbett (1991) reported a study on recurrent events of infections in 38 kidney disease patients, who use a portable dialysis machine. Two times to recurrence of an infection (days since catheter placement for each episode) were recorded as T_1 and T_2 for each patient; δ_1 and δ_2 were also recorded as the event (infection or censoring) indicators. The data contained a number of covariates including gender. In our illustration, we are interested in determining if gender has an effect on the within-subject variability of the kidney disease infection times.

The range of the event times T_j and the corresponding censoring rates (by gender and overall) are presented in Table 3. We note that McGilchrist and Aisbett (1991) analyzed this data using a Cox type model with a subject-specific frailty term. Alternatively, an accelerated failure time model (see, e.g., Fan and Datta 2011 and the references therein) with repeated measures can also be fit using the inverse probability of censoring reweighting approach. One would however like to ensure that the model errors are homogeneous in the two gender groups. Note that it amounts to testing the equality of the distribution of $|\log(T_2^*) - \log(T_1^*)|$ in the two groups.

Next, we show that, through a suitable reformulation of the problem, the test developed in this paper can be applied to test this hypothesis. To that end, consider a (hypothetical) system where state entry and exit take place at times $X^* = \log(T_1^*) \wedge \log(T_2^*)$ and $V^* = \log(T_1^*) \vee \log(T_2^*)$, respectively. Then the state sojourn time equals $W^* = |\log(T_2^*) - \log(T_1^*)|$. Furthermore, we can compute the censored version of X^*, V^* and W^* and the corresponding event indicators ξ and δ using the available data. Overall, 23 of the 38 sojourn times were fully observed. Breaking them by gen-

Table 3 Summary of kidney infection times (in days)

	T_1	Percent censored	T_2	Percent censored
Male	[2,562]	20	[7,152]	0
Female	[5,536]	32.1	[5,511]	25
Overall	[2,562]	28.9	[5,511]	18.4

der, we found that about 20 % of all sojourn times were missing for the male patients and 46 % for the female patients. For these data, the statistic T defined in Sect. 4 turned out to be 0.487 with a null standard error of 0.119. Using a two-sided Z test, we obtain a p value of 0.915 and conclude that there is no evidence to suggest that the error distributions in the two gender groups are different.

We also inspected the censoring patterns by gender in this artificial staged system. A formal test (Peto and Peto modification of the Gehan–Wilcoxon test; see, Harrington and Fleming 1982) did not indicate a difference in the censoring distribution in the two groups (p value $= 0.56$). Given the small sample size and the total amount of censoring, the study is likely to be underpowered to detect any difference in censoring distribution in the two groups even if that were the case. Fortunately, the validity of our test does not rest on this assumption. For comparison, we also performed a naive Mann–Whitney tests on the censored sojourn times W_{ij}. The resulting p value was about 0.40 leading to the same conclusion as our test.

7 Discussion

Traditional approaches of analyzing event time data include semi-parametric regression models, such as the proportional hazards model, which lead to appropriate estimating equations in the presence of right-censored data. A two-sample comparison amounts to testing the effect of a single binary covariate on event times and the test that arises from a Cox model is the log-rank test. However, in many applications, time is measured since an initiating event rather than the calendar time. In other words, we may be dealing with a sojourn time and the standard estimating equations do not hold since the censoring will not be independent of the sojourn times (Wang and Wells 1998). Furthermore, for some samples, even the state entry time may be censored. Adaptation of semi-parametric methods, such as the Cox regression to sojourn (or gap) times, have been considered through appropriate reweightings in an estimating equation (Huang 2002; Schaubel and Cai 2004; Strawderman 2005, etc.). A relatively complicated nonparametric testing methodology of comparing distributions of sojourn times was developed in Lin and Ying (2001), where the stage entry times lay below a threshold. The greatest advantage our method offers over these approaches is that it is completely nonparametric and therefore provides a valid and robust inference in the most general setting.

The methodology developed in this paper is based on a novel reweighting scheme that extends the notion of a Mann–Whitney statistic to the present setup, in which we are capable of directly comparing pairs of observable sojourn times plus additional pairs where one of the sojourn times may be missing since the exit time is right censored. The resulting statistic has desirable large sample properties including a closed form variance estimate. The large sample inference is also fairly effective in moderate samples as shown by the simulation studies. The test may be extended to a multigroup comparison in a number of standard ways, such as considering suitable linear combinations, taking a maxima, or by a quadratic form of pair-specific test scores. Another multivariate extension will be to compare two or more groups based on several sojourn times.

We have obtained a closed form estimate of the asymptotic variance of our statistics through large sample calculations (e.g., asymptotic linearity). However, the resulting summands are fairly complex and involve estimation of some auxiliary functions. As a result, care must be taken to compute it efficiently. An alternative approach will be to use the bootstrap method to estimate the variances. The proper bootstrap method in this case will be to resample the entire vector of observed data $(X_{ij},\ \xi_{ij},\ V_{ij},\ \delta_{ij})$ to obtain a bootstrap sample. A bootstrap variance estimate is then given by the empirical variance of the statistics calculated across (independently) repeated bootstrap samples.

A Mann–Whitney test is generally applied when the two distributions under comparison only differ in location. However, the theoretical reason for this convention (or assumption) arose purely from the point of view of the power function, since the Mann–Whitney test is locally the most powerful rank test in such models with logistic distributions. However, the situation is too complex in the case of censored data and a simple interpretation like this does not hold. However, it is still a valid test asymptotically from the perspective of maintaining the right size under the null hypothesis of equality of two distributions.

The nonparametric methodology of this paper is for marginal comparisons of two sojourn times even if covariates are present and are observed. It is possible to incorporate the effect of covariates that may affect the censoring distribution by recalculating the K function, which will now be individual specific and not just group specific. Of course, the asymptotic variance of the test statistic will depend on the type of models used for the censoring hazard. Resampling is a viable alternative in such cases. If one is interested in comparing the two distributions after adjusting for covariates, one can potentially use our Mann–Whitney test based on the model residuals after fitting an accelerated failure time model to the two sets of sojourn times. Of course, once again, the asymptotic variance will have to be recalculated. Another (nonparametric) way to deal with this for a low-dimensional covariate X will be to calculate a form of a conditional U-statistic given the covariate via similar inverse probability of censoring reweighting combined with smoothing techniques. An overall test may be computed from these local (i.e., for each x) test statistics by a suitable L_1 or L_2 averaging, or by taking the supremum over a suitable range of x. The details may be pursued elsewhere.

Appendix A: Technical details

Regularity conditions for the theorems

(i) $n_j/(n_1+n_2) \to c_j \in (0,1)$, for $j = 1, 2$.
(ii) $\int \omega_1^2(t)\lambda_{C_1}(t)\mathrm{d}t < \infty$ and $\int\{\omega_2^2(t)+\omega_3^2(t)\}\lambda_{C_2}(t)\mathrm{d}t < \infty$.
(iii) $\int \frac{S_2^2(w_1)}{K_1^2(v_1-)}\mathrm{d}\overline{n}_1(w_1,v_1) < \infty$, $\int \frac{F_1^2(w_2-)}{K_2^2(v_2-)}\mathrm{d}\overline{n}_2(w_2,v_2) < \infty$, and $\int \frac{I(w_1<w_2)}{K_2^2(w_1+x_2-)}\,\mathrm{d}F_1(w_1)\mathrm{d}\overline{n}_3(w_2,x_2) < \infty$.

Condition (i) is a standard design condition on the relative sample sizes in the two group settings, which helps us identify the asymptotic variance. It also ensures that we continue to have enough samples from both groups, as the total sample size increases.

Conditions (ii) ensure that certain martingales corresponding to the censoring process are squared integrable, so that an appropriate central limit theorem applies. Conditions in (iii) are technical conditions that ensure that certain summands are square integrable, so that L_2 projection calculations are possible. In particular, they also suggest that the values of the censoring survival functions K_j should not be too small on the range of V.

A.1 Proof of Theorem 1

Express

$$\begin{aligned}\sqrt{n}\,(\widehat{U}_1-\theta) &= \sqrt{n}\,(U_1-\theta)\\ &- \frac{\sqrt{n}}{n_1n_2}\sum_{i_1,i_2}\frac{I(W_{i_1,1}<W_{i_2,2})\delta_{i_1,1}\delta_{i_2,2}}{\widehat{K}_1(V_{i_1,1}-)}\left\{\frac{\widehat{K}_2(V_{i_2,2}-)-K_2(V_{i_2,2}-)}{\widehat{K}_2(V_{i_2,2}-)K_2(V_{i_2,2}-)}\right\}\\ &- \frac{\sqrt{n}}{n_1n_2}\sum_{i_1,i_2}\frac{I(W_{i_1,1}<W_{i_2,2})\delta_{i_1,1}\delta_{i_2,2}}{K_2(V_{i_2,2}-)}\left\{\frac{\widehat{K}_1(V_{i_1,1}-)-K_1(V_{i_1,1}-)}{\widehat{K}_1(V_{i_1,1}-)K_1(V_{i_1,1}-)}\right\}.\end{aligned}$$

By an L_1 analysis of the difference as in Datta et al. (2010), we can replace $\widehat{K}_j,\ j=1,\ 2$, by their in probability limits K_j in the denominator of the last two terms, provided we add an extra $o_p(1)$ term. Since

$$\sqrt{n_j}\,(\widehat{K}_j-K_j) = -\sqrt{n_j}\,K_j(\widehat{\Lambda}_j^c-\Lambda_j^c)\ +o_p(1), \tag{9}$$

by the delta method, where Λ_j^c is the cumulative censoring hazard in group j and $\widehat{\Lambda}_j^c$ is its Nelson–Aalen estimator,

$$\begin{aligned}\sqrt{n}(\widehat{U}_1-\theta) = \sqrt{n}(U_1-\theta) + \frac{\sqrt{n}}{n_1n_2}\sum_{i_1,i_2}\Bigg[&\frac{I(W_{i_1,1}<W_{i_2,2})\delta_{i_1,1}\delta_{i_2,2}}{K_1(V_{i_1,1}-)K_2(V_{i_2,2}-)}\\ &\times\left\{\widehat{\Lambda}_1^c(V_{i_1,1}-)-\Lambda_1^c(V_{i_1,1}-)+\widehat{\Lambda}_2^c(V_{i_2,2}-)-\Lambda_2^c(V_{i_2,2}-)\right\}\Bigg]+o_p(1);\end{aligned}$$

the details can be worked out by an L_2 analysis of the error term in (9). By L_2 projection calculations, as in Hoeffding's decomposition (Hoeffding 1948; Serfling 1980, page 188), the above equals

$$\begin{aligned}\frac{1}{\sqrt{n}}\Bigg[&\frac{1}{c_1}\sum_{i_1=1}^{n_1}\left\{\frac{S_2(W_{i_1,1})\delta_{i_1,1}}{K_1(V_{i_1,1}-)}-\theta\right\}+\frac{1}{c_2}\sum_{i_2=1}^{n_2}\left\{\frac{F_1(W_{i_2,2}-)\delta_{i_2,2}}{K_2(V_{i_2,2}-)}-\theta\right\}\\ &+\frac{1}{c_1}\sum_{i_1=1}^{n_1}\frac{S_2(W_{i_1,1})\delta_{i_1,1}}{K_1(V_{i_1,1}-)}\{\widehat{\Lambda}_1^c(V_{i_1,1}-)-\Lambda_1^c(V_{i_1,1}-)\}\\ &+\frac{1}{c_2}\sum_{i_2=1}^{n_2}\frac{F_1(W_{i_2,2}-)\delta_{i_2,2}}{K_2(V_{i_2,2}-)}\{\widehat{\Lambda}_2^c(V_{i_2,2}-)-\Lambda_2^c(V_{i_2,2}-)\}\Bigg]+o_p(1).\end{aligned}$$

Using a martingale representation (Andersen et al. 1993, page 178) for $\widehat{\Lambda}_j^c(\cdot) - \Lambda_j(\cdot)$, we see that the above expression equals

$$\frac{1}{\sqrt{n}}\Bigg[\frac{1}{c_1}\sum_{i_1=1}^{n_1}\Big\{\frac{S_2(W_{i_1,1})\delta_{i_1,1}}{K_1(V_{i_1,1}-)} - \theta\Big\} + \frac{1}{c_2}\sum_{i_2=1}^{n_2}\Big\{\frac{F_1(W_{i_2,2}-)\delta_{i_2,2}}{K_2(V_{i_2,2}-)} - \theta\Big\}$$
$$+\frac{1}{c_1\sqrt{n_1}}\sum_{i_1=1}^{n_1}\frac{S_2(W_{i_1,1})\delta_{i_1,1}}{K_1(V_{i_1,1}-)}\Big\{\int_0^{V_{i_1,1}-}\frac{\mathrm{d}\overline{M}_1^c(s)}{y_1(s)}\Big\}$$
$$+\frac{1}{c_2\sqrt{n_2}}\sum_{i_2=1}^{n_2}\frac{F_1(W_{i_2,2}-)\delta_{i_2,2}}{K_2(V_{i_2,2}-)}\Big\{\int_0^{V_{i_2,2}-}\frac{\mathrm{d}\overline{M}_2^c(s)}{y_2(s)}\Big\}\Bigg] + o_p(1); \tag{10}$$

here, $M_{i,j}^c(t) = N_{i,j}^c(t) - \int_0^t Y_{i,j}(u)\mathrm{d}\Lambda_j^c(u)$, $N_{i,j}^c(t) = I(V_{i,j} \le t, \delta_{i,j} = 0)$, $Y_{i,j}(t) = I(V_{i,j} \ge t)$, $\overline{M}_j = n_j^{-1/2}\sum_{i=1}^{n_j} M_{i,j}^c$ and $y_j(t) = EY_{i,j}(t)$, $j = 1, 2$. From the asymptotically linear representation of a U-statistic (Serfling 1980, page 188), the second term in the RHS of (10) equals

$$\frac{\sqrt{n_1}}{c_1}\int\frac{S_2(w_1)}{K_1(v_1-)}\Big\{\int_0^{v_1-}\frac{\mathrm{d}\overline{M}_1^c(s)}{y_1(s)}\Big\}\mathrm{d}\overline{n}_1(w_1, v_1) + o_p(\sqrt{n_1}),$$

which further equals

$$\frac{1}{c_1}\sum_{i_1=1}^{n_1}\int_0^\infty\Big\{\frac{1}{y_1(s)}\int I(v_1 > s)\frac{S_2(w_1)}{K_1(v_1-)}\mathrm{d}\overline{n}_1(w_1, v_1)\Big\}\mathrm{d}M_{i_1,1}^c(s) + o_p(\sqrt{n_1}),$$
$$= \frac{1}{c_1}\sum_{i_1=1}^{n_1}\int_0^\infty\omega_1(s)\mathrm{d}M_{i_1,1}^c(s) + o_p(\sqrt{n_1})$$

by Fubini's theorem, where ω_1 is given in (2). The third term can be handled the same way leading to the following linearization

$$\sqrt{n}(\widehat{U}_1 - \theta) = \frac{1}{\sqrt{n}}\Bigg[\frac{1}{c_1}\sum_{i_1=1}^{n_1}\Big\{\frac{S_2(W_{i_1,1})\delta_{i_1,1}}{K_1(V_{i_1,1}-)} - \theta + \int_0^\infty\omega_1(s)\mathrm{d}M_{i_1,1}^c(s)\Big\}$$
$$+\frac{1}{c_2}\sum_{i_2=1}^{n_2}\Big\{\frac{F_1(W_{i_2,2}-)\delta_{i_2,2}}{K_2(V_{i_2,2}-)} - \theta + \int_0^\infty\omega_2(s)\mathrm{d}M_{i_2,2}^c(s)\Big\}\Bigg] + o_p(1), \tag{11}$$

where ω_2 is given by (3). Therefore, as $n \to \infty$, we have

$$\sqrt{n}(\widehat{U}_1 - \theta) \xrightarrow{d} N(0, \sigma_1^2),$$

where σ_1^2 is as in the statement of Theorem 1. This proves the first assertion of Theorem 1. The linearizations for $\widehat{U}_2$ can be carried out in a similar fashion.

A.2 Estimation of variance

We estimate the asymptotic variance by the empirical variance of the linear approximation (11). Note that

$$\int \omega_j(s)\mathrm{d}\widehat{M}^c_{i,j}(s) = \omega_j(V_{i,j})\overline{\delta}_{i,j} - \int \frac{\omega_j(s)I(V_{i,j} \geq s)}{Y_j(s)}\mathrm{d}N^c_j(s),$$

where $\widehat{M}^c_{i,j}(t) = N^c_{i,j}(t) - \int_0^t Y_{i,j}(u)\mathrm{d}\widehat{\Lambda}^c_j(u)$, $\widehat{\Lambda}^c_j$ being the Nelson–Aalen estimator of Λ^c_j,

$$\begin{aligned} &= \omega_j(V_{i,j})\overline{\delta}_{i,j} - \int_0^\infty \frac{\omega_j(s)\, I(V_{i,j} \geq s)}{Y_j(s)}\mathrm{d}\left\{\sum_{i_1=1}^{n_j} N^c_{i_1,j}(s)\right\}, \\ &= \omega_j(V_{i,j})\overline{\delta}_{i,j} - \sum_{i_1=1}^{n_j} \frac{\omega_j(V_{i_1,j})\, I(V_{i,j} \geq V_{i_1,j})\overline{\delta}_{i_1,j}}{Y_j(V_{i_1,j})}. \end{aligned}$$

This justifies the choice of the summands $S_{1,i}$. The other parts can be obtained in a similar fashion. Estimation of ω_j uses the principle of inverse probability of censoring reweighting (Datta et al. 2010). Consistency of $\widehat{\sigma}^2_j$ can be established using projection techniques for generalized U-statistics and laws of large number results for reweighting.

A.3 Proof of Theorem 2

Asymptotic linearization of the test statistic is obtained as a linear combination of the linear approximations of the statistics $\widehat{U}(1,2)$ and $\widehat{U}(2,1)$ as obtained under Theorem 1.

Acknowledgments This research was supported in part by a grant from the United States National Science Foundation (DMS-0706965). We thank an anonymous reviewer and an associate editor for their very helpful comments.

References

Andersen, P. K., Borgan, O., Gill, R. D., Keiding, N. (1993). *Statistical models based on counting processes*. New York: Springer.

Datta, S., Bandyopadhyay, D., Satten, G. A. (2010). Inverse probability of censoring weighted U-statistics for right censored data with applications. *Scandinavian Journal of Statistics*, *37*, 680–700.

Fan, J., Datta, S. (2011). Fitting accelerated failure time models to clustered survival data with potentially informative cluster size. *Computational Statistics & Data Analysis*, *55*, 3295–3303.

Harrington, D. P., Fleming, T. R. (1982). A class of rank test procedures for censored survival data. *Biometrika*, *69*, 553–566.

Hoeffding, W. (1948). A class of statistics with asymptotically normal distribution. *Annals of Mathematical Statistics*, *19*, 293–325.

Huang, Y. (2002). Censored regression with the multistate accelerated sojourn times model. *Journal of the Royal Statistical Society, Series B*, *64*, 17–29.
Latta, R. B. (1977). Generalized Wilcoxon statistics for the two-sample problem with censored data. *Biometrika*, *64*, 633–635.
Lin, D. Y., Ying, Z. (2001). Nonparametric tests for the gap time distributions of serial events based on censored data. *Biometrics*, *57*, 369–375.
Mann, H. B., Whitney, D. R. (1947). On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics*, *18*, 50–60.
McGilchrist, C. A., Aisbett, C. W. (1991). Regression with frailty in survival analysis. *Biometrics*, *47*, 461–466.
Prentice, R. L. (1978). Linear rank tests with right censored data. *Biometrika*, *65*, 167–179.
Satten, G. A., Datta, S. (2002). Marginal estimation for multistage models: waiting time distributions and competing analyses. *Statistics in Medicine*, *21*, 3–19.
Schaubel, D. E., Cai, J. (2004). Regression methods for gap time hazard functions of sequentially ordered multivariate failure time data. *Biometrika*, *91*, 291–303.
Schisterman, E., Rotnizky, A. (2001). Estimation of the mean of a K-sample U-statistic with missing outcomes and auxiliaries. *Biometrika*, *88*, 713–725.
Serfling, R. J. (1980). *Approximation theorems of mathematical statistics*. New York: Wiley.
Strawderman, R. L. (2005). The accelerated gap times model. *Biometrika*, *92*, 647–666.
Wang, W., Wells, M. T. (1998). Nonparametric estimation of successive duration times under dependent censoring. *Biometrika*, *85*, 561–572.
Wilcoxon, F. (1945). Individual comparisons by ranking methods. *Biometrics*, *1*, 80–83.