Published in final edited form as:

Lifetime Data Anal. 2017 April ; 23(2): 207–222. doi:10.1007/s10985-015-9347-7.

Nonparametric inference for the joint distribution of recurrent marked variables and recurrent survival time

Laura M. Yee¹ and Kwun Chuen Gary Chan²

¹Center for Devices and Radiological Health, U.S. Food and Drug Administration, 10903 New Hampshire Ave, Silver Spring, MD 20993, USA

²Department of Biostatistics, University of Washington, Campus Box 357232, Seattle, WA 98195-7232, USA

Abstract

Time between recurrent medical events may be correlated with the cost incurred at each event. As a result, it may be of interest to describe the relationship between recurrent events and recurrent medical costs by estimating a joint distribution. In this paper, we propose a nonparametric estimator for the joint distribution of recurrent events and recurrent medical costs in right-censored data. We also derive the asymptotic variance of our estimator, a test for equality of recurrent marker distributions, and present simulation studies to demonstrate the performance of our point and variance estimators. Our estimator is shown to perform well for a wide range of levels of correlation, demonstrating that our estimators can be employed in a variety of situations when the correlation structure may be unknown in advance. We apply our methods to hospitalization events and their corresponding costs in the second Multicenter Automatic Defibrillator Implantation Trial (MADIT-II), which was a randomized clinical trial studying the effect of implantable cardioverter-defibrillators in preventing ventricular arrhythmia.

Keywords

Induced informative censoring; Marked variables; Medical cost data; Nonparametric estimation; Recurrent events

1 Introduction

Cost data are being collected more frequently in order to assess cost-effectiveness of new treatments. In particular, more studies are collecting longitudinal cost data, where cost information is collected at random medical event times. As such data becomes more common, it may be of interest to characterize the time between recurring events, and the medical cost associated with these events. For example, the time between recurring hospitalizations might be correlated with the cost incurred at each hospitalization. In such a

Correspondence to: Laura M. Yee.

Electronic supplementary material The online version of this article (doi:10.1007/s10985-015-9347-7) contains supplementary material, which is available to authorized users.

situation, it may be of interest to inspect the joint distribution of recurrent hospitalization time and recurrent hospitalization cost.

The purpose of estimation of a bivariate distribution is typically for exploratory and descriptive analyses. One particular advantage of nonparametric estimation is that the results are not based on additional parametric/semiparametric modeling assumptions. The results of a bivariate estimate can be presented in a table, and the results are useful for actuarial applications in health insurance. Notably, the bivariate distribution of recurrent events and costs is the fundamental building block of the Sparre-Andersen risk process (Rolski et al. 2009). Consistent estimates of the bivariate distribution can be used to determine insurance premiums and to ensure sufficient reserve funds to meet claim obligations.

Medical cost is known as a marked variable. A marked variable is only observable at survival events which are subject to censoring. Marked variables are therefore substantially different from variables that can be collected at baseline. When lifetime cost is of interest, we say that lifetime cost is a mark of survival. This means that lifetime cost is only observed at an event time when a patient has not been censored. In the presence of right censoring, a complication called induced informative censoring occurs on the cost scale (Glasziou et al. 1990; Lin et al. 1997; Huang and Louis 1998; Strawderman 2000; Bang and Tsiatis 2000). For example, a patient who is quite sick accumulates medical cost more quickly on average than a healthy patient at any given time, and therefore his or her cumulative cost would tend to be higher than that of his or her healthy counterpart at both the censoring time and the survival time, even when the censoring and survival times are independent.

Time between events that generate medical cost may be observed in a recurrent fashion. After enrollment due to an initial event (such as hospitalization due to myocardial infarction), events (such as subsequent hospitalization) incurring medical cost will recur periodically throughout the course of study. Recurrent time data also has the complication of informative censoring (Wang and Chang 1999; Pena et al. 2001). In this situation, a patient who is ill will have more hospitalizations, and therefore the time between his or her last recurrence time and censoring time is likely to be shorter. In addition, recurrent cost will be a marked variable for recurrent hospitalizations.

In previous work, Huang and Louis (1998) presented a nonparametric estimator for the joint distribution of non-recurrent survival time and marked variables. Huang and Louis used a survival analysis formulation to define a cumulative mark specific hazard which did not rely upon an assumption of non-informative censoring on the mark scale. In addition, they derived the asymptotic variance of their estimator using a martingale approach. We plan to extend the method of Huang and Louis to account for recurrent survival time and recurrent marked variables.

Focusing on recurrent event times, several authors have discussed nonparametric estimation for the recurrent survival distribution, which have been discussed in a recent review paper by Zhu (2013). In particular, Wang and Chang (1999) used a survival analysis approach to define a recurrent cumulative hazard function, which incorporated mass and risk sets which were weighted within an individual's recurrent events before weighting across individuals.

Their method allowed for within-person correlation amongst recurrent survival times, which handles informative censoring of recurrent events. The approach proposed by Wang and Chang has been extended to estimate the distribution of alternating recurrent events (Huang and Wang 2005). In addition, Luo and Huang (2011) proposed a resampling method in order to analyze recurrent events, which was shown to be equivalent to the method of Wang and Chang when the number of resamples tends to infinity. We will extend the method of Wang and Chang, along with the work of Huang and Louis, to construct an estimator for the joint distribution of recurrent survival time and marked variables.

In this paper, we aim to present a nonparametric estimator for the joint distribution of recurrent survival time and marked variables. We will also present its asymptotic variance, a test for equality of bivariate recurrence distributions, simulation results, and an example based on implantation of cardioverter-defibrillators in the second Multicenter Automatic Defibrillator Implantation Trial (MADIT-II).

2 Methods

2.1 Notation

For individual i , $i = 1, \dots, n$, we record recurrent events j , with the time of the initial event set to 0. Define T_{ij} to be recurrent event times. That is, T_{ij} denotes the gap time between the $(j-1)$ th event to the j th event for the i th individual. Let C_i be the censoring time for individual i , and let M_{ij} be the observed marked variable, which is observed if the j th event is observed.

Adopting notation similar to that of Wang and Chang (1999), let k_i be the index of the last observed recurrence time that is subject to censoring, such that for person i :

$$\sum_{j=1}^{k_i-1} T_{ij} \leq C_i \quad \text{and} \quad \sum_{j=1}^{k_i} T_{ij} > C_i.$$

Since we only observe marked variables and recurrent events before C_i , we define Y_{ij} to be the observed recurrent event times as follows:

$$Y_{ij} = \begin{cases} T_{ij} & \text{if } j = 1, \dots, k_i - 1 \\ T_{ik_i}^+ & \text{if } j = k_i \end{cases}$$

where $T_{ik_i}^+$ is the time from the $(j-1)$ th event to C_i . We additionally make two assumptions:

1. There is an unobserved subject-level random variable X such that for all individuals i , the recurrent event times and marked random variables (T_{i1}, M_{i1}) , (T_{i2}, M_{i2}) , ... are independent and identically distributed given X_i .
2. The censoring time, C_i is independent of $(X_i, \{T_{i1}, M_{i1}\}, \{T_{i2}, M_{i2}\}, \dots)$.

The first assumption therefore allows for dependence of the recurrent data within a subject, and the second assumption is an independent censoring condition. We note that X is an unobserved subject-level latent variable that is used to characterize within-subject correlations. The above assumptions are similar to the assumptions A1 and A2 in Wang and Chang (1999), which are much weaker than the independent and identically distributed recurrent events as in Pena et al. (2001). The subject-level latent variable X is unobservable and therefore plays no role in the estimators in Sect. 2.2; this assumption is quite weak and allows the recurrent marker processes to be correlated within the same individual. The introduction of X is similar to the use of random effects to characterize correlated outcomes within an individual in a longitudinal study. An important difference is that we do not require normality or any parametric assumptions of the random effect distribution. This innovation is made in Wang and Chang (1999).

2.2 Estimation of the joint marked recurrent distribution

Let us first define the parameters that we will employ in the definition of our joint distribution. Let $S(t)$ be the recurrent survival function, $S(t) = P(T_{ij} > t)$ and let $G(t, m) = P(T_{ij} \leq t, M_{ij} \leq m)$ be the joint distribution of the recurrent events and the marked variables. Once again following the definitions of Wang and Chang (1999), define $H(t) = E[I(T_{i1} \leq t)I(C_i \leq t)]$. Now, let $\Lambda(t, m)$ be the joint cumulative hazard for recurrent events and the marked variable. As noted by Huang and Louis (1998), $\Lambda(t, m)$ can be written as follows:

$$\Lambda(t, m) = \int_{[0, t]} \frac{F(ds, u)}{H(s)}, \quad (1)$$

where $F(t, m) = E[I(T_{i1} \leq t, M_{i1} \leq m)I(T_{i1} \leq C_i)]$ is the cumulative joint sub-distribution function of recurrent events and recurrent marked variables amongst uncensored participants.

Then the goal is to estimate the joint distribution of the recurrent event times and the marked variable, $G(t, m)$:

$$G(t, m) = P(T_{ij} \leq t, M_{ij} \leq m) = \int_0^t S(s)\Lambda(ds, m). \quad (2)$$

To estimate, $G(t, m) = \int_0^t S(s)\Lambda(ds, m)$, we will first focus on estimating $\Lambda(t, m)$ of Eq. (1) under a recurrent event setting. As defined by Wang and Chang (1999), an estimator for $H(t)$, the weighted mean risk set, is defined as follows:

$$\hat{H}(t) = \frac{1}{n} \sum_{i=1}^n \left[\frac{1}{k_i^*} \sum_{j=1}^{k_i^*} I(y_{ij} \geq t) \right],$$

where $k_1^* = \max(k_i - 1, 1)$. Now, extending the work of Wang and Chang (1999) to include marked variables, we let

$$\hat{F}(t, m) = \frac{1}{n_i} \sum_{i=1}^n \left[\frac{I(k_i \geq 2)}{k_i^*} \sum_{j=1}^{k_i^*} I(y_{ij} \leq t, m_{ij} \leq m) \right]$$

be the mass at recurrent event time t and marked variable value m , first averaging over a person's recurrent events. Similarly to Wang and Chang (1999), we exclude the last recurrent event time due to its bias. To illustrate this idea, consider a hypothetical example in which the recurrent event times take two possible values $t = 1$ or 2 with equal probabilities. Let the censoring time, which is distributed independent of the recurrent events, be uniform $(0, 10)$ distributed. Then the probability that the censoring time lies within intervals with $t = 2$ is twice than with $t = 1$. Therefore, there will be disproportionately more $t = 2$ for the last and censored recurrent event. On the other hand, uncensored event times within an individual are identically distributed. In particular, they are distributed as the first recurrent event time. Therefore, we can replace the first recurrent event as in Huang and Louis (1998) by the average of all uncensored events.

Note that $\hat{F}(t, m)$ is an extension of the marked counting process of Huang and Louis, which is an extension of the failure counting process commonly used in survival analysis. In those counting processes, only uncensored events are counted. In recurrent events, the counting process also counts only uncensored events, which is here denoted by the indicator $I(k_i \geq 2)$. In contrast, $\hat{H}(t)$ is analogous to the at-risk processes which counts both uncensored and censored events, and therefore the $I(k_i \geq 2)$ is not included. When there is only a single survival event, our definition of $\hat{F}(t, m)$ and $\hat{H}(t)$ reduces to those in Huang and Louis (1998).

Given estimates $\hat{H}(t)$ and $\hat{F}(t, m)$, an estimator for $\hat{\Lambda}(t, m)$ can be defined as follows:

$$\hat{\Lambda}(t, m) = \int_{[0, t]} \frac{\hat{F}(ds, m)}{\hat{H}(s)}.$$

To estimate $G(t, m)$, we also need to estimate $S(t)$. Note that $\hat{\Lambda}(t, \infty) = \hat{\Lambda}(t)$, where $\hat{\Lambda}(t)$ is the cumulative hazard estimator of recurrent event times proposed by Wang and Chang (1999):

$$\hat{\Lambda}(t) = \int_{[0, t]} \frac{\hat{F}(ds)}{\hat{H}(s)},$$

where $\hat{F}(t) = \hat{F}(t, \infty)$. Now, we let $\hat{S}(t)$ be the estimated recurrent survival function by Wang and Chang (1999), where $\hat{S}(t) = \prod_{[0, t]} [1 - \hat{\Lambda}(ds)]$. Finally, we propose to estimate $G(t, m)$ by plugging in $\hat{\Lambda}(t, m)$ for $\Lambda(t, m)$, and $\hat{\Lambda}(t)$ for $\Lambda(t)$ in the product-limit estimator of $S(t)$. Then our estimator for the joint distribution of recurrent survival time and recurrent marked variables, $\hat{G}(t, m)$, is written as follows:

$$\hat{G}(t, m) = \int_{[0, t]} \prod_{[0, s)} [1 - \hat{\Lambda}(ds)] \hat{\Lambda}(ds, m).$$

Note that $\hat{G}(t, \infty)$ reduces exactly to $1 - \hat{S}(t)$. When there is only a single survival event, $\hat{S}(t)$ reduces to the Kaplan–Meier estimator. Furthermore, in this case, $\hat{\Lambda}(t, m)$ and $\hat{G}(t, m)$ reduce to the corresponding estimators proposed by Huang and Louis (1998).

The estimated recurrent survival distribution, $\hat{S}(t)$, is identifiable on $[0, \tau]$, where $\tau = \sup\{t : S(t) > 0\}$ can be thought of as the end of study follow-up. As was discussed by Huang and Louis (1998), although we are not able to identify the full survival distribution, the joint distribution, $\hat{G}(t, m)$, is identifiable for the region $[0, \tau] \times \mathbb{R}$.

2.3 Variance estimation

We shall now discuss the asymptotic *i.i.d.* representation of $\hat{G}(t, m) - G(t, m)$ and the resulting estimator for the variance of $\hat{G}(t, m)$. In the Appendix, we use the functional delta method (Vaart 1998) to show that there exists a function $\eta_i(t, m)$ such that

$$\begin{aligned} (\hat{G}(t, m) - G(t, m)) &= \frac{1}{n} \sum_{i=1}^n \eta_i(t, m) + o_p(1/\sqrt{n}), \quad \text{and} \\ \eta_i(t, m) &= \frac{1}{n} \sum_{i=1}^n \left[\int_0^t S(s-) \phi_i(ds, m) + \int_0^t \phi_i^{WC}(s) \Lambda(ds, m) \right] + o_p\left(\frac{1}{\sqrt{n}}\right). \end{aligned}$$

In the above equations, $\phi(t, m)$ and $\phi_i^{WC}(t)$ are influence functions of $(\hat{\Lambda}(t, m) - \Lambda(t, m))$ and $(\hat{S}(t) - S(t))$, respectively. The formulations of $\phi_i(t, m)$ and $\phi_i^{WC}(t)$ are also included in the Appendix. In addition, a consistent estimator for the variance of $\hat{G}(t, m)$ can be written as follows:

$$\begin{aligned} \widehat{\text{Var}}[\hat{G}(t, m)] &= \frac{1}{n} \sum_{i=1}^n \left[\int_0^t \hat{S}(s-) \hat{\phi}_i(ds, m) + \int_0^t \hat{\phi}_i^{WC}(s) \hat{\Lambda}(ds, m) \right]^2 \\ &= \frac{1}{n} \sum_{i=1}^n [\hat{\eta}_i(t, m)]^2, \end{aligned}$$

where $\hat{\eta}_i(t, m)$, $\hat{\phi}_i(t, m)$, and $\hat{\phi}_i^{WC}(t)$ are also given in the Appendix.

2.4 Equal distribution assumption

The proposed methodology depends on the assumption that events have the same distribution across recurrences. To evaluate this assumption, we propose a test of equal distribution. Define the test statistic $T = \sup |\hat{G}(t, m) - \hat{G}_{HL}(t, m)|$, where $\hat{G}_{HL}(t, m)$ is the estimator proposed by Huang and Louis (1998) using only data corresponding to the first recurrence. Then we can find a critical value, b , against which we can compare the test statistic, T . The null distribution of T and the critical value can be found in the following manner via multiplier bootstrap, which follows a procedure similar to that of Lin (1997):

1. Estimate the empirical influence functions for $\hat{G}(t, m)$ and $\hat{G}_{HL}(t, m)$ using the estimators defined in Sect. 2.3 for $i = 1, \dots, n$. Let the estimators for these influence functions be denoted by $\hat{\eta}_i(t, m)$ and $\hat{\eta}_i^{HL}(t, m)$. We substitute unknown functions with the proposed estimators based on all recurrences in the calculation of both $\hat{\eta}_i(t, m)$ and $\hat{\eta}_i^{HL}(t, m)$ since the proposed estimators are consistent to the truth under the null hypothesis.
2. Generate standard normal distributed multipliers, W_{ih} for each observation independent of the data.
3. Calculate $b_h = \sup \left| 1/n(\sum_i W_{ih} \hat{\eta}_i(t, m) - \sum_i W_{ih} \hat{\eta}_i^{HL}(t, m)) \right|$.
4. Repeat steps 2 and 3 for H independent samples of multipliers, and then obtain b , the 95th percentile of $b_1, \dots, b_H$.

After calculating the critical value b , we reject the null hypothesis of equal distribution if $T > b$. If we reject the null hypothesis, we determine that there is evidence that the distribution differs across recurrences. In this situation, it may be appropriate to apply other methods. See Sect. 5 for a related discussion.

The validity of the above procedure is due to the fact that conditioned on observed data, the only randomness in b_h comes from the simulated multipliers. Following the arguments in Lin (1997), the asymptotic distributions of $n^{-1/2} \sum_i W_{ih} \hat{\eta}_i(t, m)$ and $n^{-1/2} \sum_i W_{ih} \hat{\eta}_i^{HL}(t, m)$ conditioned to data are equivalent to the asymptotic distributions of $\sqrt{n}G(t, m)$ and $\sqrt{n}G_{HL}(t, m)$ under the null hypothesis. Therefore, $b_1, \dots, b_H$ approximates the null distribution of T .

3 Simulation

In this section, we use numerical simulations to illustrate the performance of the proposed methods in Sect. 2. We employed simulation parameters similar to the simulation experiments conducted by Huang and Louis (1998) and Wang and Chang (1999). Simulation results were based on 1,000 replications.

We generated $n = 100$ or $n = 200$ Beta(3,1)-distributed frailty values x_i , $i = 1, \dots, n$, with probability distribution function $f(x) = 3x^2$. Next, correlated sets of recurrent survival times and marked variables were generated using Frank's bivariate family, a copula model (Genest 1987). To generate data according to this model, we first sample independent (u_{ij}, v_{ij}) , where $i = 1, \dots, n$ and $j = 1, \dots$ are standard uniform distributed. Next, we set $R_{ij} = \alpha^{u_{ij}} + (\alpha - \alpha^{u_{ij}})v_{ij}$, where $\alpha = \exp(a)$. The constant a allows for different levels of correlation between t_{ij} and m_{ij} . Finally, we defined $t_{ij} = \left(-\frac{1}{x_i} \log(1 - u_{ij}) \right)^{1/2}$ and $m_{ij}^* = \log_\alpha [R_{ij} / (R_{ij} + (1 - \alpha)v_{ij})]$. This resulted in recurrence times, t_{ij} , that were Weibull distributed with the survival function $\exp(-x_i \sqrt{t})$, which was a survival setup similar to Wang and Chang (1999). Finally, we transform the standard uniform distributed m_{ij}^* to generate costs, m_{ij} which were gamma($10 \times$

x_i ; rate = $0.01 \times x_i$) distributed. Marginally, m_{ij} is distributed with mean 1000 and variance 150000. Following the work of Wang and Chang (1999), we let $C_i = 2$ for all individuals. With these simulation parameters, the mean k_i was 2.6. In our simulations, we let a vary between -10 , 0 , and 10 , which corresponds to Kendall's rank correlations between t_{ij} and m_{ij} of 0.67 , 0 , and -0.67 .

A summary of the simulation results can be found in Table 1. We summarize at values of t and m , where t is equal to 1.0 or 2.0 , and m takes values in 1000 , 1250 , 1500 , and ∞ . Reported values $t = 1.0$ and $t = 2.0$ were chosen to be representative of the marginal distribution of T_{ij} , and correspond to cumulative probabilities of $G(t = 1.0, m = \infty) = 0.51$ and $G(t = 2.0, m = \infty) = 0.93$. Similarly, $m = 1000$, 1250 , and 1500 correspond to marginal probabilities of $G(t = \infty, m = 1000) = 0.55$, $G(t = \infty, m = 1250) = 0.78$, and $G(t = \infty, m = 1500) = 0.90$. For each pair (t, m) , we present three rows of information: (1) the true value followed by the empirical variance of the estimators (in parentheses); (2) our estimated value followed by our estimated variance (in parentheses); and (3) the empirical coverage of the 95% confidence intervals. From the simulation results, we see that the bias of the estimator is small in all situations. The estimated variance is also close to the empirical variance for all t and m , confirming the validity of our derived asymptotic variance formula. We also performed simulations for $n = 50$. The results are similar and are shown in Online Resource 1. Additional simulations confirmed that the empirical type I errors of the hypothesis testing method are close to the nominal values.

4 Second Multicenter Automatic Defibrillator Implantation Trial (MADIT-II)

We illustrate our methods using data from the second Multicenter Automatic Defibrillator Implantation Trial (MADIT-II) (Moss et al. 2002). MADIT-II was a clinical trial that aimed to prevent sudden cardiac death in patients at high risk for cardiac arrhythmia. Participants were randomized to receive an implantable cardioverter-defibrillator (ICD) or standard treatment. Cost and healthcare utilization data were also collected, and were reported by Zwanziger et al. (2006). Hospitalization costs are the marked variables of interest. Medical cost is used as an example in many of the existing recurrent or non-recurrent marked variable literature. For example, see Lin (2000), Liu et al. (2008), Huang (2009), Chan and Wang (2010) and Fang et al. (2011), among others.

Conducted from July 11, 1997 to January 16, 2002, participants were considered for inclusion in the study if they had a previous myocardial infarction and had a left ventricular ejection fraction less than 0.30 . Investigators enrolled 1232 participants from 76 hospitals in the US and Europe. Zwanziger and others reported cost-effectiveness results for 664 ICD participants and 431 standard care participants (Zwanziger et al. 2006). This cost study further excluded participants from European hospitals and participants with large amounts of missing medical cost data. The majority of censoring events were due to administrative censoring. In the ICD group, 79% of participants were administratively censored at the end of study follow-up, while 85% were administratively censored in the standard group. We demonstrate our methods using the 1095 participants and data presented by Zwanziger et al. Further information about the study has been documented by the original investigators (Moss et al. 2002; Zwanziger et al. 2006).

In order to mimic hospitalization costs, equal consecutive daily costs were summed and any resulting cost events less than \$500 were omitted. The data we employed has many small costs with frequent and deterministic recurrent events, which are likely to be pharmacy or outpatient visits. One limitation of the dataset is that there is no indication of the type of service provided. Therefore, for illustration, we chose a cut value in advance, which was meant to mimic hospitalization costs. In addition, the start time in the ICD group was set to the day after each participant's ICD surgery was completed, since we are interested in studying the joint distribution of recurrent medical cost and survival time after ICD implantation.

Participants were followed for up to 4.5 years. Observed follow-up time was 1.8 years (SD=1.08 years) and 1.7 years (SD=1.05 years), on average, for the ICD group and the standard treatment group. On the person level, we observed a median 2.0 (IQR 1.0–5.0) and 1.0 (IQR 0–3.5) number of recurrent hospitalizations amongst ICD participants and standard treatment participants, respectively.

Table 2 summarizes the results of our analysis of the MADIT-II data. The estimate of the joint distribution and the associated standard deviation (in parentheses) are reported for a variety of recurrent survival times and recurrent costs ($t=1, 2, 3$ years and $m=\$1000, \$2000, \$7000, \100000). In addition, we compare our proposed joint distribution (left) to the method of Huang and Louis (1998) applied only to the first recurrent event (right). Although the outcomes in the data used for the two methods are different, they are comparable when the equal recurrent marker distribution assumption holds. Using the proposed test in Sect. 2.4, we obtained a p value of 0.58 for the standard group and a p value of 0.22 for the ICD group, each obtained using 1000 multiplier samples. The null hypothesis of equal distributions is not rejected at the 5% significance level.

Referring to the results of our proposed method (labeled “All Recurrences”), we see that standard care participants have relatively few low recurrent cost events for any length of recurrent survival time. We see that after ICD implantation, participants in the ICD group tended to have a higher proportion of low recurrent costs and more frequent recurrent times. In other words, for the ICD patients, we observe a joint distribution that is more concentrated in the region of low recurrent costs and shorter recurrence times. Now we turn to the results employing the estimator of Huang and Louis (1998) (labeled “First Recurrence”), which only uses the first recurrent hospitalization time and first recurrent medical cost. The point estimates in this case are similar to the point estimates of our proposed method, but as expected, the estimated standard deviations are larger since information on subsequent recurrences are not utilized. Compared to the method of Huang and Louis (1998), our estimator reduces up to 33% of the estimated standard deviation in this example.

To examine the association between the recurrence times and costs, Fig. 1 displays conditional distribution functions of costs corresponding to shorter and longer recurrences. In the Figure, *solid curves* are estimates for $P(M \leq m | T \leq 1)$ and *dashed curves* are estimates for $P(M \leq m | T > 1)$. It is interesting to note that the cost distribution is not associated with whether recurrence times are shorter or longer than one year for standard treatment, but

costs are negatively associated with recurrence times for ICD. Recurrences that are longer than one year have a higher chance to have small hospitalization costs for ICD patients.

5 Concluding remarks

This paper discussed a method to estimate the joint distribution between recurrent survival times and recurrent marked variables. We presented a nonparametric estimator and a variance estimator, proposed a test for equality of recurrent marker distributions and illustrated our methods via simulation and a data example.

An associate editor raised a question of whether other bivariate survival distribution estimators such as the Dabrowska (1988) estimator can be extended to the case of recurrent marker processes. We would like to clarify that even without recurrent events, the joint distribution of the survival event and marked variables can be estimated by the Huang-Louis estimator, but not other bivariate estimators such as the Dabrowska estimator. This subtlety is due to how the marked variables are defined and when they are observed. When the focus is on biomarkers observable at baseline, Shi et al. (2014) considered the Dabrowska estimator for the joint distribution. In Huang and Louis (1998), the marked variables are observable at the failure event but not at baseline, and the Dabrowska estimator is not applicable. Because medical costs are associated with recurrent hospitalization events which are not defined at baseline, the setting in Huang and Louis (1998) is more suitable and the Dabrowska estimator cannot be applied.

Our proposed method assumes the same joint distribution across recurrences. In our analysis of the MADIT-II data, we saw that estimates across all recurrences were similar to those using only the first recurrence, indicating that this assumption was not violated. We also proposed a test for the equality of distributions. However, this identical distribution assumption may not be tenable in some situations, especially when recurrences may be viewed as progression of disease (Wang and Chen 2000). In this case, the marginal distribution of second and subsequent recurrent events cannot be identified nonparametrically, and a semiparametric modeling approach will be needed. As such, future work may extend this framework to a more general setting such as multivariate recurrences.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank Dr. Arthur Moss and Dr. Hongwei Zhao for access to the MADIT-II data. The authors would also like to thank the editor, an associate editor and three reviewers for their constructive comments that greatly improved the paper. Kwun Chuen Gary Chan is partially supported by US National Institutes of Health Grant R01 HL 122212.

Appendix

We wish to find the influence function of $\hat{G}(t, m)$. To do so, we will first discuss the uniform consistency of $\hat{G}(t, m)$, then find the influence function of $\hat{A}(t, m)$, and finally use

those results to get the final influence function of interest. We use the same notation and setup of Sect. 2 of the paper.

Uniform consistency of $\hat{\Lambda}(t, m)$ and $\hat{G}(t, m)$

Note that $\hat{F}(t, m)$ and $\hat{H}(t)$ are empirical processes, $E(\hat{F}(t, m)) = F(t, m)$ and $E(\hat{H}(t)) = H(t)$. Therefore, $\hat{F}(t, m)$ and $\hat{H}(t)$ are uniformly consistent estimators for $F(t, m)$ and $H(t)$ by the Glivenko–Cantelli theorem (Vaart 1998). Note that $\hat{\Lambda}(t, m)$ is a functional of $\hat{F}(t, m)$ and $\hat{H}$, and therefore, $\hat{\Lambda}(t, m)$ is uniformly consistent to $\Lambda(t, m)$ by Lemma A.1 in Lin et al. (2000). Also, $\hat{S}(t)$ is uniformly consistent according to Wang and Chang (1999). Since $\hat{G}(t, m)$ is a functional of $\hat{S}(t)$ and $\Lambda(t, m)$, it follows that $\hat{G}(t, m)$ is uniformly consistent to $G(t, m)$ by Lemma A.1 in Lin et al. (2000).

Influence function: $\hat{\Lambda}(t, m)$

First, note that $\hat{\Lambda}(t, m) = \int_{[0, t]} \hat{F}(ds, m) / \hat{H}(s)$ depends on the pair $(\hat{F}(t, m), \hat{H}(t))$ through two maps:

$$(A, B) \rightarrow \left(A, \frac{1}{B}\right) \rightarrow \int_0^t \frac{1}{B} dA.$$

Then, the functional derivative of the maps at (F, H) is:

$$(\alpha, \beta) \rightarrow \left(\alpha, -\frac{\beta}{H^2}\right) \rightarrow \int_0^t \frac{d\alpha}{H} - \int_0^t \frac{\beta dF}{H^2},$$

where $\alpha = \hat{F} - F$ and $\beta = \hat{H} - H$. Now, by the functional delta method (Vaart 1998) and simplification we have:

$$\begin{aligned}
 \hat{\Lambda}(t, m) - \Lambda(t, m) &= \int_0^t \frac{1}{H(s)} [\hat{F}(ds, m) - F(ds, m)] \\
 &\quad - \int_0^t \frac{\hat{H}(s) - H(s)}{(H(s))^2} F(ds, m) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
 &= \int_0^t \frac{\hat{F}(ds, m)}{H(s)} - \Lambda(t, m) - \int_0^t \frac{\hat{H}(s)F(ds, m)}{(H(s))^2} \\
 &\quad + \Lambda(t, m) + o_p\left(\frac{1}{\sqrt{n}}\right) \\
 &= \frac{1}{n} \sum_{i=1}^n \left[\frac{I(k_i \geq 2)}{k_i^*} \sum_{j=1}^{k_i^*} \left(I(y_{ij} \leq t, m_{ij} \leq m) \frac{1}{H(y_{ij})} \right) \right. \\
 &\quad \left. - \int_0^t \frac{F(ds, m)}{(H(s))^2} \frac{1}{k_i^*} \sum_{j=1}^{k_i^*} I(y_{ij} \geq s) \right] \\
 &\quad + o_p\left(\frac{1}{\sqrt{n}}\right) \\
 &= \frac{1}{n} \sum_{i=1}^n \phi_i(t, m) + o_p\left(\frac{1}{\sqrt{n}}\right)
 \end{aligned}$$

where

$$\begin{aligned}
 \phi_i(t, m) &= \frac{I(k_i \geq 2)}{k_i^*} \sum_{j=1}^{k_i^*} \left(I(y_{ij} \leq t, m_{ij} \leq m) \frac{1}{H(y_{ij})} \right) \\
 &\quad - \int_0^t \frac{F(ds, m)}{(H(s))^2} \frac{1}{k_i^*} \sum_{j=1}^{k_i^*} I(y_{ij} \geq s)
 \end{aligned}$$

is the influence function of $\hat{\Lambda}(t, m) - \Lambda(t, m)$. Then a consistent estimator for $\phi_i(t, m)$, $\hat{\phi}_i(t, m)$, can be calculated by plugging in $\hat{H}(t)$ and $\hat{F}(t, m)$ for $H(t)$ and $F(t, m)$.

Influence function: $\hat{G}(t, m)$

Now, we wish to find the influence function for $\hat{G}(t, m)$, the estimated joint distribution of recurrent survival time and marked variable, where $\hat{G}(t, m) = \int_0^t \hat{S}(s -) \hat{\Lambda}(ds, m)$. The influence function for the recurrent survival function, $\hat{S}(t)$, as presented in Wang and Chang, is written as follows:

$$\begin{aligned}
 \phi_i^{WC}(t) &= S(t) \left[\int_0^t \frac{1}{H^2(u)} \left(\frac{1}{k_i^*} \sum_{j=1}^{k_i^*} I(y_{ij} \geq u) \right) F(du) \right. \\
 &\quad \left. - \frac{I(k_i \geq 2)}{k_i^*} \sum_{j=1}^{k_i^*} \frac{I(y_{ij} < t)}{H(y_{ij})} \right].
 \end{aligned}$$

Now, completing steps similar to those in the previous section of the Appendix, we use both $\phi_i(t, m)$ and $\phi_i^{WC}(t)$ to obtain the influence function for $G(t, m)$:

$$\begin{aligned}\hat{G}(t, m) - G(t, m) &= \int_0^t S(s-) [\hat{\Lambda}(ds, m) - \Lambda(ds, m)] \\ &\quad + \int_0^t [\hat{S}(s-) - S(s-)] \Lambda(ds, m) + o_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \left[\int_0^t S(s-) \phi_i(ds, m) + \int_0^t \phi_i^{WC}(s) \Lambda(ds, m) \right] \\ &\quad + o_p\left(\frac{1}{\sqrt{n}}\right) \\ &= \frac{1}{n} \sum_{i=1}^n \eta_i(t, m) + o_p\left(\frac{1}{\sqrt{n}}\right).\end{aligned}$$

Then an estimator for $\eta_i(t, m)$, $\hat{\eta}_i(t, m)$, can be estimated by plugging in $\hat{S}(t)$, $\hat{\phi}_i(t, m)$, $\hat{\phi}_i^{WC}(t)$, and $\hat{\Lambda}(t, m)$ for $S(t)$, $\phi_i(t, m)$, $\phi_i^{WC}(t)$, and $\Lambda(t, m)$.

References

- Bang H, Tsiatis AA. Estimating medical costs with censored data. *Biometrika*. 2000; 87(2):329–343.
- Chan KCG, Wang MC. Backward estimation of stochastic processes with failure events as time origins. *Ann Appl Stat*. 2010; 4(3):1602–1620. [PubMed: 21359167]
- Dabrowska DM. Kaplan-Meier estimate on the plane. *Ann Stat*. 1988; 16(4):1475–1489.
- Fang HB, Wang J, Deng D, Tang ML. Estimating the mean of a mark variable under right censoring on the basis of a state function. *Comput Stat Data Anal*. 2011; 55(4):1726–1735.
- Genest C. Frank's family of bivariate distributions. *Biometrika*. 1987; 74(3):549–555.
- Glasziou PP, Simes RJ, Gelber RD. Quality adjusted survival analysis. *Stat Med*. 1990; 9(11):1259–1276. [PubMed: 2277877]
- Huang Y. Cost analysis with censored data. *Med Care*. 2009; 47(7):S115–S119. [PubMed: 19536024]
- Huang Y, Louis TA. Nonparametric estimation of the joint distribution of survival time and mark variables. *Biometrika*. 1998; 85(4):785–798.
- Huang CY, Wang MC. Nonparametric estimation of the bivariate recurrence time distribution. *Biometrics*. 2005; 61(2):392–402. [PubMed: 16011685]
- Lin DY. Non-parametric inference for cumulative incidence functions in competing risks studies. *Stat Med*. 1997; 16(8):901–910. [PubMed: 9160487]
- Lin DY. Proportional means regression for censored medical costs. *Biometrics*. 2000; 56(3):775–778. [PubMed: 10985215]
- Luo X, Huang CY. Analysis of recurrent gap time data using the weighted risk-set method and the modified within-cluster resampling method. *Stat Med*. 2011; 30(4):301–311. [PubMed: 20963733]
- Lin DY, Feuer EJ, Etzioni R, Wax Y. Estimating medical costs from incomplete follow-up data. *Biometrics*. 1997; 53(2):419–434. [PubMed: 9192444]
- Lin DY, Wei LJ, Yang I, Ying Z. Semiparametric regression for the mean and rate functions of recurrent events. *J R Stat Soc Ser B*. 2000; 62(4):711–730.
- Liu L, Huang X, O'Quigley J. Analysis of longitudinal data in the presence of informative observational times and a dependent terminal event, with application to medical cost data. *Biometrics*. 2008; 64(3):950–958. [PubMed: 18162110]
- Moss AJ, Zareba W, Hall WJ, Klein H, Wilber DJ, Cannom DS, Daubert JP, Higgins SL, Brown MW, Andrews ML. Prophylactic implantation of a defibrillator in patients with myocardial infarction and reduced ejection fraction. *New Engl J Med*. 2002; 346(12):877–883. [PubMed: 11907286]

- Pena EA, Strawderman RL, Hollander M. Nonparametric estimation with recurrent event data. *J Am Stat Assoc.* 2001; 96(456):1299–1315.
- Rolski, T., Schmidli, H., Schmidt, V., Teugels, J. *Stochastic processes for insurance and finance.* Wiley; NewYork: 2009.
- Shi H, Cheng Y, Li J. Assessing diagnostic accuracy improvement for survival or competing-risk censored outcomes. *Can J Stat.* 2014; 42(1):109–125.
- Strawderman RL. Estimating the mean of an increasing stochastic process at a censored stopping time. *J Am Stat Assoc.* 2000; 95(452):1192–1208.
- van der Vaart, AW. *Asymptotic statistics.* Cambridge University Press; New York: 1998.
- Wang MC, Chang SH. Nonparametric estimation of a recurrent survival function. *J Am Stat Assoc.* 1999; 94(445):146–153. [PubMed: 24244058]
- Wang MC, Chen YQ. Nonparametric and semiparametric trend analysis for stratified recurrence times. *Biometrics.* 2000; 56(3):789–794. [PubMed: 10985217]
- Zhu H. Non-parametric analysis of gap times for multiple event data: an overview. *Int Stat Rev.* 2013; 82(1):106–122.
- Zwanziger J, Hall WJ, Dick AW, Zhao H, Mushlin AI, Hahn RM, Wang H, Andrews ML, Mooney C, Wang H, J MA. The cost effectiveness of implantable cardioverter-defibrillators: results from the Multicenter Automatic Defibrillator Implantation Trial (MADIT)-II. *J Am coll Cardiol.* 2006; 47(11):2310–2318. [PubMed: 16750701]

Fig. 1. Conditional cumulative distributions of medical costs for the MADIT-II trial. *Solid lines* represent the conditional distributions with recurrence up to 1 year, and *dashed lines* represent the conditional distributions with survival beyond 1 year. The x-axis is given in logarithmic scale

Table 1

Simulation results for the joint distribution of recurrent events and recurrent marked variables for each time ($t = 1$ or 2) and marked category ($m = 1000, 1250, 1500, \infty$): row 1—true values (empirical variance), row 2—empirical means of the estimator (empirical means of the variance estimator), and row 3—empirical coverage at the 95% nominal level

	$m = 1000$	$m = 1250$	$m = 1500$	$m = \text{Inf}$
$a = -10$				
N = 100				
$t = 1$	0.453 (0.0019)	0.51 (0.002)	0.517 (0.002)	0.518 (0.002)
	0.449 (0.0017)	0.507 (0.0019)	0.514 (0.0019)	0.515 (0.0019)
	0.94	0.94	0.94	0.94
$t = 2$	0.547 (0.002)	0.76 (0.0016)	0.867 (0.0011)	0.928 (0.0006)
	0.545 (0.0018)	0.758 (0.0016)	0.866 (0.001)	0.928 (0.0006)
	0.93	0.94	0.93	0.93
N = 200				
$t = 1$	0.453 (0.0008)	0.51 (0.0009)	0.517 (0.0009)	0.518 (0.0009)
	0.452 (0.0009)	0.51 (0.001)	0.517 (0.001)	0.519 (0.001)
	0.95	0.96	0.96	0.96
$t = 2$	0.547 (0.0009)	0.76 (0.0008)	0.867 (0.0006)	0.928 (0.0004)
	0.546 (0.0009)	0.759 (0.0008)	0.867 (0.0005)	0.929 (0.0003)
	0.95	0.94	0.94	0.92
$a = 0$				
N = 100				
$t = 1$	0.284 (0.0014)	0.403 (0.0018)	0.47 (0.0019)	0.518 (0.002)
	0.284 (0.0013)	0.401 (0.0016)	0.467 (0.0018)	0.515 (0.0019)
	0.94	0.93	0.95	0.94
$t = 2$	0.51 (0.0019)	0.72 (0.0015)	0.838 (0.0011)	0.928 (0.0006)
	0.51 (0.0017)	0.72 (0.0014)	0.838 (0.001)	0.928 (0.0006)
	0.94	0.93	0.94	0.93
N = 200				
$t = 1$	0.284 (0.0007)	0.403 (0.0008)	0.47 (0.0009)	0.518 (0.0009)
	0.284 (0.0007)	0.403 (0.0008)	0.47 (0.0009)	0.519 (0.001)
	0.95	0.96	0.96	0.96
$t = 2$	0.51 (0.0009)	0.72 (0.0008)	0.838 (0.0006)	0.928 (0.0004)
	0.508 (0.0009)	0.718 (0.0008)	0.838 (0.0005)	0.929 (0.0003)
	0.95	0.95	0.93	0.92
$a = 10$				
N = 100				
$t = 1$	0.118 (0.0007)	0.302 (0.0015)	0.422 (0.0019)	0.517 (0.002)
	0.119 (0.0007)	0.302 (0.0014)	0.421 (0.0018)	0.515 (0.0019)
	0.94	0.94	0.94	0.94

	$m = 1000$	$m = 1250$	$m = 1500$	$m = \text{Inf}$
$t = 2$	0.48 (0.0016)	0.702 (0.0014)	0.829 (0.001)	0.928 (0.0006)
	0.483 (0.0017)	0.704 (0.0013)	0.83 (0.001)	0.928 (0.0006)
	0.96	0.94	0.94	0.93
N = 200				
$t = 1$	0.118 (0.0003)	0.302 (0.0007)	0.422 (0.0009)	0.517 (0.0009)
	0.118 (0.0003)	0.304 (0.0007)	0.424 (0.0009)	0.519 (0.001)
	0.95	0.96	0.95	0.96
$t = 2$	0.48 (0.0009)	0.702 (0.0007)	0.829 (0.0005)	0.928 (0.0004)
	0.479 (0.0009)	0.703 (0.0007)	0.83 (0.0005)	0.929 (0.0003)
	0.96	0.94	0.95	0.92

Table 2
Results from the second Multicenter Automatic Defibrillator Implantation Trial (MADIT-II)

Years	Recurrent cost (\$, thousands)				First recurrence (H&L)			
	All recurrences							
	1	2	7	100	1	2	7	100
Standard								
1	0.148 (0.017)	0.277 (0.02)	0.471 (0.022)	0.616 (0.019)	0.161 (0.025)	0.275 (0.03)	0.45 (0.03)	0.598 (0.024)
2	0.177 (0.025)	0.329 (0.029)	0.579 (0.034)	0.756 (0.028)	0.194 (0.034)	0.334 (0.04)	0.568 (0.044)	0.76 (0.035)
3	0.209 (0.042)	0.368 (0.046)	0.623 (0.052)	0.812 (0.053)	0.226 (0.049)	0.38 (0.061)	0.621 (0.068)	0.83 (0.073)
ICD								
1	0.268 (0.025)	0.365 (0.025)	0.542 (0.023)	0.714 (0.014)	0.252 (0.029)	0.336 (0.032)	0.521 (0.028)	0.687 (0.017)
2	0.321 (0.041)	0.438 (0.048)	0.639 (0.054)	0.824 (0.056)	0.311 (0.047)	0.427 (0.058)	0.639 (0.063)	0.819 (0.062)
3	0.366 (0.087)	0.486 (0.096)	0.694 (0.108)	0.887 (0.124)	0.367 (0.111)	0.492 (0.13)	0.717 (0.149)	0.91 (0.185)

Proposed estimation of the joint distribution of recurrent hospitalization time (at $t = 1, 2$, or 3 years) and recurrent medical cost (at \$1,000, \$2,000, \$7,000, or \$100,000) for the proposed method (left) or the method of Huang and Louis applied only to the first recurrence (right)