Published in final edited form as:

J Am Stat Assoc. 2017 ; 112(518): 826–836. doi:10.1080/01621459.2016.1180988.

Nonparametric Benefit–Risk Assessment Using Marker Process in the Presence of a Terminal Event

Yifei Sun^a, Chiung-Yu Huang^b, and Mei-Cheng Wang^a

^aDepartment of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD

^bSidney Kimmel Comprehensive Cancer Center, Johns Hopkins University, Baltimore, MD

Abstract

Benefit–risk assessment is a crucial step in medical decision process. In many biomedical studies, both longitudinal marker measurements and time to a terminal event serve as important endpoints for benefit–risk assessment. The effect of an intervention or a treatment on the longitudinal marker process, however, can be in conflict with its effect on the time to the terminal event. Thus, questions arise on how to evaluate treatment effects based on the two endpoints, for the purpose of deciding on which treatment is most likely to benefit the patients. In this article, we present a unified framework for benefit–risk assessment using the observed longitudinal markers and time to event data. We propose a cumulative weighted marker process to synthesize information from the two endpoints, and use its mean function at a prespecified time point as a benefit–risk summary measure. We consider nonparametric estimation of the summary measure under two scenarios: (i) the longitudinal marker is measured intermittently during the study period, and (ii) the value of the longitudinal marker is observed throughout the entire follow-up period. The large-sample properties of the estimators are derived and compared. Simulation studies and data examples exhibit that the proposed methods are easy to implement and reliable for practical use. Supplemental materials for this article are available online.

Keywords

Kernel smoothing; Longitudinal marker process; Multiple event process; Survival analysis

1. Introduction

Assessing benefits and risks is a crucial step in medical decision making process. The purpose of benefit–risk assessment is to determine whether the benefits of an intervention or treatment outweigh its risks based on a given measure. In many clinical trials or biomedical

CONTACT: Yifei Sun, ysun@jhu.edu, Department of Biostatistics, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD 21205.

Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

Supplementary Materials

Tables, Figures and Detailed Proofs: Additional simulation study results; Plots in the analysis of CPCRA ddi/ddC trial; Proof of Theorem 3.1, 3.2, 4.1, and Lemma 1, 2 and 3. (Supplementary.pdf)

studies, a conventional way of risk assessment is to analyze the time to an event of interest. Statistical methods such as the log-rank test and Cox's proportional hazards model are widely used for risk assessment based on the event time. On the other hand, longitudinally measured patient-centered outcomes or biomarkers are also frequently collected, because they characterize patients' health status and quality of life over time. For example, in the Didanosine/Zalcitabine trial conducted by Terry Bein Community Programs for Clinical Research on AIDS (CPCRA) (Abrams et al. 1994), time to AIDS progression or death is the primary endpoint; moreover, the Karnofsky score, which quantifies patients' general well-being and physical quality of life, is assigned by study investigators at each follow-up visit. The longitudinal marker measurements, such as quality of life score, offer insights into patients' experience and perceptions and serve as important endpoints in evaluating a treatment. Thus, a question arises on how to assess benefits and risks based on both time to event and longitudinal marker, for the purpose of deciding on which treatment is most likely to benefit the patients.

Because the longitudinal measurements and time to the event are often correlated in nature, the occurrence of the terminal event can induce informative drop-out to the collection of longitudinal markers. Thus, a conventional longitudinal data analysis which fails to account for the correlated terminal event can result in biased estimation. In the literature, many authors have proposed to employ a joint model of the longitudinal marker process and the terminal event time process to make valid inference. For example, Wu and Carroll (1988), Tsiatis, Degruetola, and Wulfsohn (1995), and Hogan and Laird (1997) linked the two outcome processes via subject-specific random effects, while Henderson, Diggle, and Dobson (2000), Wang and Taylor (2001), and Xu and Zeger (2001) considered using a time-varying latent process to link the two processes. Although the joint modeling approach is appropriate for describing treatment effects on the longitudinal marker and the time to the terminal event separately, it may be inadequate for decision-making. If a treatment has favorable effects on both endpoints, the decision is straightforward; however, if one treatment shows an advantage on survivorship but a disadvantage on longitudinal marker, then the decision is more difficult to make. In the latter scenario, a summary measure that integrates information from event time and longitudinal marker is desired, and decision can be made by comparing the summary measure across different treatments.

Quality-adjusted survival analysis (Gelber, Gelman, and Goldhirsch 1989; Glasziou, Simes, and Gelber 1990) is a useful tool that incorporates survival time and quality of life (QoL) into a summary measure. By weighting the durations of different health states by their respective utility values, a single endpoint is constructed to summarize the duration of survival and the QoL. Nonparametric estimation of the expected quality-adjusted survival time has been studied by many authors, including (Huang and Louis 1999; Shen, Pulkstenis, and Hoseyni 1999; Zhao and Tsiatis 1999; Murray and Cole 2000). Ideally, the multiple health-state models require clearly defined states that reflect severity of disease progression over time. The use of the multiple health-state model may not be appropriate when the transitions of states are unclear or loosely defined. In such cases, utilizing the QoL measurements in quality-adjusted survival analysis could be analytically preferable. In particular, Hwang, Tsauo, and Wang (1996) and Glasziou et al. (1998) considered using QoL measures over time as the dynamic weight, instead of assigning a fixed weight to a

specific health state. In Hwang, Tsauo, and Wang (1996), in addition to a cohort study from which the survival function of the time to the terminal event is readily estimable, another cross-sectional survey needs to be conducted to estimate the quality-of-life weight. The validity of the estimator then relies on the assumption that the subjects in the cross-sectional survey must be a random sample from the original cohort study population. Glasziou et al. (1998) handled the missingness of QoL measure between intermittent measurements by interpolation, which requires rather strict assumptions on the individual or population mean trajectory of QoL. To ensure an accurate decision-making process, it is desirable to develop standardized and validated methodologies for studying quality-adjusted survival.

In the context of synthesizing information from both the terminal event process and the longitudinal marker process, we develop a unified framework for benefit-risk assessment to facilitate decision-making. A summary measure integrating the two outcomes is proposed, including the expected quality-adjusted survival time as a special case. We consider different nonparametric approaches for the summary measure under two scenarios: (i) the longitudinal marker is measured intermittently until terminal event or loss to follow-up, and (ii) the value of the longitudinal marker is observed throughout the entire follow-up period. In contrast with the approach of Hwang, Tsauo, and Wang (1996), which requires an extra cross-sectional survey, our estimators for the proposed summary measure are self-contained in the sense that the estimates can be obtained by using data within one study. The contents of this article are organized as follows: In Section 2, we define cumulative weighted marker process and use its mean function at a prespecified time point as a summary measure for benefit-risk assessment. In Section 3, we consider nonparametric estimation of the mean function of the cumulative weighted marker process when the longitudinal marker is intermittently observed. In Section 4, a nonparametric estimator of the mean function is proposed when the longitudinal marker is continuously observed. In Section 5, we report the results of simulation studies. In Section 6, the proposed methods are applied to two clinical trials for illustration. Discussions on generalizations of our methods are given in Section 7.

2. Cumulative Weighted Marker Process and Benefit-Risk Assessment

Let $\{Y(t), t \geq 0\}$ be a longitudinal marker process, where $Y(t)$ is a nonnegative marker measurement at time t . Denote the time to the terminal event of interest by D , where D is possibly correlated with marker process $Y(\cdot)$. Here, we consider benefit-risk assessment based on the time to the terminal event and the longitudinal marker process before the terminal event, that is, $\{Y(t), D, 0 \leq t \leq D\}$, as the value of $Y(\cdot)$ after D is either not defined or not of interest. For ease of discussion, we assume that a larger marker value indicates a more favorable result throughout this article. Define the *cumulative weighted marker process*

$$M(t) = \int_0^t w(u)Y(u)I(D \geq u)du,$$

where $w(u)$ is a prespecified weight function and $Y(u)I(D > u)$ is a marker process that takes the value 0 after the terminal event. When setting $w(\cdot) = 1$, $M(t)$ is the area under the marker trajectory before the time point t or the terminal event, whichever occurs first. Note that $M(t)$ can be viewed as an endpoint that integrates information from both the longitudinal marker

process and the survival time. An ideal treatment or intervention should prolong survival while maintaining higher marker values over time, thus leading to a large value of $M(t)$ at any time point.

Taking expectation of $M(t)$, we define the cumulative mean function

$$\mu(t) = E\{M(t)\} = \int_0^t w(u)E\{Y(u)I(D \geq u)\}du.$$

This gives the area under curve for the weighted mean function $w(u)E\{Y(u)I(D \geq u)\}$. The weight function $w(\cdot)$ can be set to reflect the clinical importance of outcomes at different time points, and a detailed discussion on selection of $w(\cdot)$ is in Remark 2.1. In the special case where $w(\cdot) = 1$ and $Y(\cdot) = 1$, $\mu(t)$ reduces to the restricted mean survival time up to t

(Irwin 1949); moreover, in the absence of the terminal event, $\mu(t) = \int_0^t E\{Y(u)\}du$ is the area

under the expected marker trajectory up to t (Sun and Wu 2003). We propose to use $\mu(\tau)$, the cumulative mean function at a prespecified time point τ , as a benefit-risk summary measure, where a treatment with higher value of $\mu(\tau)$ is preferred. In what follows, we consider two scenarios to illustrate the use of the proposed summary measure.

Example 2.1 (QoL and survival)

With advances in treatment and supportive care, treatment decision-making for patients with advanced cancer are increasingly complex. Because cure is elusive, it is broadly recognized that prolonging survival may not be the only goal of treatment and that maintaining QoL could also be an important task, as patients may not be willing to trade lower QoL for longer survival. Let D be the time to death and $Y(t)$ be the QoL measurement at t . The cumulative weighted marker process $M(t)$ integrates QoL and survival into a composite outcome for clinical decision analysis. For the special case where $w(\cdot) = 1$ and $Y(\cdot)$ is a step function with discrete values, $\mu(\tau)$ is the mean quality-adjusted survival time restricted to time τ under multiple health states (Gelber, Gelman, and Goldhirsch 1989; Glasziou, Simes, and Gelber 1990). As an important assumption of the multiple health-state model, the quality adjusted duration of a health state needs to be proportional to the actual duration of that health state. The assumption is most clearly satisfied if the QoL measure is constant over the entire duration of each health state. Therefore, when QoL measure varies considerably within one health state or the transitions between health states are unclear, progressive health states are generally difficult to define in relation to QoL data. For example, in tumors or degenerative cases, such as muscular dystrophy, the common mode of QoL is in steady decline, which is hard to convert to a multiple health-state model. The proposed summary measure $\mu(\tau)$ directly incorporates the longitudinal QoL measure collected in clinical trials and does not rely on the proportionality assumption. Comparison based on $\mu(\tau)$ can assist investigators to evaluate trade-offs between survival and QoL.

Example 2.2 (Multiple events and survival)

In many longitudinal studies, the occurrence of multiple events are commonly encountered and serve as important endpoints. For the Beta-Blocker Evaluation of Survival Trial (The

Beta-Blocker Evaluation of Survival Trial Investigators 2001), an advanced chronic heart failure clinical trial, in addition to overall survival, which is the primary endpoint of the study, clinical outcomes such as hospitalization, myocardial infarction, and heart transplantation are also of interest. We denote by T_1 , T_2 , and T_3 , the time of the three secondary endpoints, and by D the time to death. To incorporate information from the multiple event process, one can define $Y(t) = \sum_{i=1}^3 I(T_i \geq t) + 1$. Then, the stochastic process $Y(\cdot)$, which decreases by 1 when any one of the three non-fatal events occurs, can be viewed as a score that reflects patients' disease burden and health condition over time. By setting $w(\cdot) = 1$, the summary measure $\mu(\tau) = E\{\sum_{i=1}^3 \min(T_i, D, \tau) + \min(D, \tau)\}$ is the expected sum of four types of event-free survival times up to τ (Claggett et al. 2014).

In the first example, the longitudinal marker process $Y(\cdot)$ is usually measured at intermittent time points, while in the second example, $Y(\cdot)$ is completely observed throughout the follow-up period. We then develop different estimating procedures corresponding to the two types of observed data in Section 3 and 4.

Of note, in the definition of the cumulative mean function, the weight $w(\cdot)$ can be chosen to reflect the importance of marker and lifetime in different time periods in benefit-risk decision making. For example, when the immediate impact of a treatment is of primary interest, one may use a decreasing weight function of time; when the long-term benefit is of more importance, one may use an increasing weight function. In cost-effectiveness analysis, even if the marker and lifetime are equally important throughout the follow-up period, health economics researchers, including Weinstein and Stason (1977); Viscusi (1995) and many others, have suggested that marker adjusted life years should be discounted to convert future health benefits into terms that can be compared with present values, so that combined analysis of costs and health benefits can be properly performed. A common approach is to discount health outcomes at an annual rate between 3% and 5%, that is, setting $w(u) = a^u$ with $0.95 \leq a \leq 0.97$ (Gravelle and Smith 2001; Walker and Kumaranayake 2002; Robberstad 2005). In practice, the weighting scheme should be specified in consultation with domain experts. The choice of the weight function will affect the power of two-sample testing, as suggested by the simulation studies in supplementary materials. We recommend to use different weight functions to gain more insight into treatment effects; for example, one may consider using $w_1(u) = a^u$ and $w_2(u) = a^{\tau-u}$ with a ranging from 0 to 1.

3. Nonparametric Estimation of $\mu(t)$ When Marker is Intermittently Observed

In this section, we consider nonparametric estimation of the cumulative mean function $\mu(t)$ ($0 \leq t \leq \tau$) in the case where the longitudinal marker process $Y(\cdot)$ is measured intermittently. In practice, the survival time D is subject to right censoring due to study end or premature dropout. We denote the censoring time by C and assume that C is independent with $\{Y(\cdot), D\}$. Define $X = \min(D, C)$ and $\delta = I(D \leq C)$. Let $N^*(\cdot)$ be the counting process for the potential data collecting times of the marker $Y(\cdot)$, where the rate function of $N^*(\cdot)$ is $\lambda^*(t)$, that is, $E\{dN^*(t)\} = \lambda^*(t) dt$, $t \geq 0$. Then, the counting process $N(t) = I(X \leq t)N^*(t)$ gives the number of observations of the marker before time t , that is, $Y(\cdot)$ is observed only at the time points where $N(\cdot)$ jumps. We further assume that $N^*(\cdot)$ is independent with $\{D, C, Y(\cdot)\}$, then

the rate function of the observation time process $N(t)$ is $\lambda(t) = S_X(t)\lambda^*(t)$ with $S_X(t) = \Pr(X \geq t)$. In other words, $\lambda(t)$ gives the instantaneous “risk” of the marker being measured at time t . In many applications, although the follow-up visits were scheduled on a regular basis, the actual observation times were irregularly spaced across subjects due to various practical issues, such as mistimed visits. We further assume that $\lambda(t) > 0$ for $0 \leq t \leq \tau$, which requires the collection of intermittently observed times from all subjects to be reasonably dense in $[0, \tau]$ as sample size becomes large. The observations $\{X_i, Y_i(t)dN_i(t), 0 \leq t \leq \tau, i = 1, \dots, n\}$ are assumed to be independent replicates of $\{X, Y(t)dN(t), 0 \leq t \leq \tau\}$.

Two major challenges lie in the estimation of $\mu(t) = E\{M(t)\}$. First, because $Y(\cdot)$ is observed at discrete time points during the course of follow-up, the cumulative weighted marker process $M(t)$ is not evaluable. Second, even in the ideal case that $Y(\cdot)$ is completely observed up to X , the induced informative censoring hampers the development of statistical methods. Although it is usually reasonable to assume that the terminal event time D and the censoring time C are independent, $M(D)$ and $M(C)$ are usually positive correlated. For example, a healthier subject may maintain a higher marker value over time, hence having larger $M(C)$ as well as $M(D)$. The naive method of treating $\{M_i(X_i), i = 1, \dots, n\}$ as right censored data and estimating the distribution of $M(D)$ using the Kaplan–Meier estimator can result in substantial bias. In what follows, we propose two consistent estimators for $\mu(t)$ and study their large-sample properties.

3.1. A Kernel Smoothing Approach

To construct a nonparametric estimator for $\mu(t) = \int_0^t w(u)E\{Y(u)I(D \geq u)\}du$, we first note that the function $\mu(t)$ can be decomposed as

$$\mu(t) = \int_0^t w(u)S_D(u)E\{Y(u)|D \geq u\}du, \quad (3.1)$$

where $S_D(u) = \Pr(D \geq u)$ is the survival function of D and $r(u) = E\{Y(u)|D \geq u\}$ is the expected marker value of survivors at u . Under independent censoring, subjects in the risk set at time u are a representative sample of event-free individuals at time u in the target population. As a result, it can be shown that $E\{Y(u)|D \geq u\} = E\{Y(u)|X \geq u\}$. We propose to estimate $r(u)$ with

$$\hat{r}_h(u) = \frac{\sum_{i=1}^n \int_0^\tau K_h(u-s)Y_i(s)I(X_i \geq s)dN_i^*(s)}{\sum_{i=1}^n \int_0^\tau K_h(u-s)I(X_i \geq s)dN_i^*(s)}, u \in [h, \tau-h], \quad (3.2)$$

where $K_h(x) = h^{-1}K(x/h)$ is a kernel function with bandwidth h , and $K(\cdot)$ satisfies

$\int_{-1}^1 K(x)dx = 1$ and $\int_{-1}^1 xK(x)dx = 0$. We define $0/0$ to be 0 in the case where the denominator

of $\hat{r}_h(u)$ is 0. It is easy to see that $\hat{r}_h$ is a locally weighted average of nearby marker values and is a natural extension of the Nadaraya–Watson estimator. If the uniform kernel is employed, that is, $K(x) = \mathbb{I}(|x| < 1)/2$, the denominator of $\hat{r}_h(u)$ is the total number of observations in the time interval $[u - h, u + h]$, while the numerator is the sum of all the observed marker value in $[u - h, u + h]$. To avoid biased estimates in the boundary region $[0, h)$ and $(\tau - h, \tau]$, we set $\hat{r}_h(u) = \hat{r}_h(h)$ for $u \in [0, h)$, and $\hat{r}_h(u) = \hat{r}_h(\tau - h)$ for $u \in (\tau - h, \tau]$. It is shown in supplementary materials that $\hat{r}_h(\cdot)$ is uniformly consistent on $[0, \tau]$.

Replacing $E\{Y(u) | D = u\}$ with $\hat{r}_h(u)$ and $S_D(u)$ with the Kaplan–Meier estimator $\hat{S}_D(u)$ in (3.1), we propose to estimate $\mu(t)$ by

$$\hat{\mu}_A(t) = \int_0^t w(u) \hat{S}_D(u) \hat{r}_h(u) du. \quad (3.3)$$

Theorem 3.1 summarizes the large-sample properties of $\hat{\mu}_A(t)$. Define

$$M_i^D(t) = N_i^D(t) - \int_0^t I(X_i \geq u) d\Lambda^D(u), \text{ where } \Lambda^D(t) \text{ is the cumulative hazard function of } D \text{ and } N_i^D(t) = I(D_i \leq t, \Delta_i = 1).$$

Theorem 3.1—Under Assumptions (A1)–(A5) in the Appendix, for $0 \leq t \leq \tau$,

$n^{1/2}\{\hat{\mu}_A(t) - \mu(t)\}$ has an asymptotically iid representation

$$n^{1/2}\{\hat{\mu}_A(t) - \mu(t)\} = n^{-1/2} \sum_{i=1}^n \Psi_i(t) + o_p(1), \text{ where}$$

$$\begin{aligned} \Psi_i(t) = & \int_0^t \frac{\mu(u) dM_i^D(u)}{S_X(u)} - \mu(t) \int_0^t \frac{dM_i^D(u)}{S_X(u)} + \int_0^t \frac{w(u) S_D(u) Y_i(u) I(X_i \geq u) dN_i^*(u)}{\lambda(u)} \\ & - \int_0^t \frac{w(u) S_D(u) r(u) I(X_i \geq u) dN_i^*(u)}{\lambda(u)}. \end{aligned}$$

Moreover, as $n \rightarrow \infty$, $\sqrt{n}\{\hat{\mu}_A(t) - \mu(t)\}$ converges weakly to a zero mean normal distribution with variance $E\{\Psi_1(t)^2\}$.

It is worthwhile to point out that the main technical challenge in proving $\sqrt{n}$ –consistency of $\hat{\mu}_A(t)$ is that the Kaplan–Meier estimator $\hat{S}_D(\cdot)$ is $\sqrt{n}$ –consistent while the kernel-type estimator $\hat{r}_h(\cdot)$ is $\sqrt{nh}$ –consistent, thus commonly used techniques such as functional delta method can not be directly applied. It is shown in the supplementary materials that, by under smoothing $r(u)$ using bandwidth $h \propto n^{-\nu}$ ($1/4 < \nu < 1/2$), $\hat{\mu}_A(t)$ can achieve $\sqrt{n}$ –consistency.

Although, as a common practice, marker values of survivors are summarized and analyzed for treatment comparison, caution should be paid when interpreting the function $r(u) = E\{Y(u) | D = u\}$, because the survivor population changes over time and may not be

representative of the originally randomized population defined at time zero. To see this, suppose D and $Y(\cdot)$ are correlated through a frailty V , where a larger value of V inflates the risk of the terminal event and decreases the value of marker process simultaneously. If a treatment decreases the risk of the terminal event but does not affect $Y(\cdot)$, it can be shown that $E(V|D = u)$ of the treatment group is larger than or equal to that of the control group at any time u . As a result, the survivors, based on which inference for $r(u)$ are drawn, are not comparable between treatment and controls as the terminal events occur along the time. In this case, $r(u)$ of treatment group may be lower than or equal to that of the control group. Hence, comparisons based on $r(u)$ may yield incorrect conclusion about the treatment effects on the longitudinal marker process.

3.2. A Computationally More Efficient Approach

In practice, numerical integration is employed to approximate the integral in (3.3). Thus, the estimated curve $\hat{r}_h(u)$ needs to be evaluated at a large number of grid points. To reduce the computational burden, we consider an alternative estimator that does not require numerical integration in evaluating the estimator. Specifically, the second estimator is motivated by the equality

$$E\{Y(u)I(X \geq t)dN^*(u)\} = E\{Y(u)|D \geq u\}\lambda(u)du,$$

which holds under the assumption that $N^*(\cdot)$ is independent of $\{Y(\cdot), D, C\}$ and C is independent of $\{Y(\cdot), D\}$. Provided $\lambda(u) > 0$ for $u \in [0, t]$, we have

$$\mu(t) = \int_0^t w(u)S_D(u)\lambda(u)^{-1}E\{Y(u)I(X \geq u)dN^*(u)\}.$$

Note that $Y(u)I(X \geq u)dN^*(u)$ is a stochastic process that takes nonzero values only at the time when $dN^*(u) > 0$, so the stochastic process is completely observed and its mean function $E\{Y(u)I(X \geq u)dN^*(u)\}$ can be consistently estimated by its empirical average $n^{-1}\sum_{i=1}^n Y_i(u)I(X_i \geq t)dN_i^*(u)$. Then a non-parametric estimator for $\mu(t)$ is given by

$$\hat{\mu}_B(t) = \frac{1}{n} \sum_{i=1}^n \int_0^t \frac{w(u)\hat{S}_D(u)Y_i(u)I(X_i \geq u)dN_i^*(u)}{\hat{\lambda}_h(u)}$$

where $\hat{\lambda}_h(u) = n^{-1}\sum_{i=1}^n \int_0^\tau K_h(u-s)dN_i(s)$ is a nonparametric smoothed estimator estimator for the rate function $\lambda(u)$. Note that $\hat{\lambda}_h(\cdot)$ can be viewed as an extension of kernel density estimator proposed by Wang and Chiang (2002). As before, we set $\hat{\lambda}_h(u) = \hat{\lambda}_h(h)$ for $u \in (0, h]$ and $\hat{\lambda}_h(u) = \hat{\lambda}_h(\tau - h)$ for $u \in [\tau - h, \tau]$ to avoid boundary effect of the kernel estimator. Theorem 3.2 summarizes the large sample properties of $\hat{\mu}_B(t)$, with proofs given in supplementary materials.

Theorem 3.2—Under the assumptions in Theorem 3.1, for $0 \leq t \leq \tau$, $n^{1/2}\{\hat{\mu}_B(t) - \mu(t)\}$ has an asymptotically iid representation $n^{1/2}\{\hat{\mu}_B(t) - \mu(t)\} = n^{-1/2} \sum_{i=1}^n \Psi_i(t) + o_p(1)$. Moreover, as $n \rightarrow \infty$, $\sqrt{n}\{\hat{\mu}_B(t) - \mu(t)\}$ converges weakly to a zero mean normal distribution with variance $E\{\Psi_1(t)^2\}$.

Interestingly, the two nonparametric estimators $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ are asymptotically equivalent. Note that the latter evaluates the smoothed function $\hat{\lambda}_h(\cdot)$ only at the time when marker values are observed, while the former evaluates the smoothed function $\hat{r}_h(\cdot)$ on a much finer grid for numerical integration. Hence, $\hat{\mu}_B(t)$ is computationally more convenient than $\hat{\mu}_A(t)$. The simulation study in Section 5.1 shows that the two estimators have similar performance with finite sample size. We then recommend the use of $\hat{\mu}_B(t)$ to estimate $\mu(t)$. For the standard error estimation, the variance $E\{\Psi_1(t)^2\}$ can be consistently estimated by $n^{-1} \sum_{i=1}^n \hat{\Psi}_i(t)^2$ under the assumptions in Theorem 3.1, and

$$\begin{aligned} \hat{\Psi}_i(t) = & \int_0^t \frac{\hat{\mu}_B(u) d\hat{M}_i^D(u)}{\hat{S}_X(u)} - \hat{\mu}_B(t) \int_0^t \frac{d\hat{M}_i^D(u)}{\hat{S}_X(u)} + \int_0^t \frac{w(u) \hat{S}_D(u) Y_i(u) I(X_i \geq u) dN_i^*(u)}{\hat{\lambda}_h(u)} \\ & - \int_0^t \frac{w(u) \hat{S}_D(u) \hat{r}_h(u) I(X_i \geq u) dN_i^*(u)}{\hat{\lambda}_h(u)}, \end{aligned}$$

where $\hat{\Lambda}^D(t) = n^{-1} \sum_{i=1}^n \int_0^t \hat{S}_X(u)^{-1} dN_i^D(u)$ is the Nelson–Aalen estimator of $\Lambda^D(t)$, and

$$\hat{M}_i^D(t) = N_i^D(t) - \int_0^t I(X_i \geq u) d\hat{\Lambda}_D(u).$$

In practice, investigators are often interested in comparing two treatment to find a more effective therapy. Suppose there are two groups, say, group 1 and group 2. With subscript j indicating the j th group, we consider testing the null hypothesis $H_0: \mu_1(\tau) = \mu_2(\tau)$ as a reference for decision-making. Let n_j be the number of subjects in the j th group, and let $\pi_j = \lim_{n \rightarrow \infty} n_j/(n_1 + n_2)$, $j = 1, 2$. Straightforward test statistics can be constructed as

$$W_A = \hat{\mu}_{A1}(\tau) - \hat{\mu}_{A2}(\tau) \text{ or } W_B = \hat{\mu}_{B1}(\tau) - \hat{\mu}_{B2}(\tau), \quad (3.4)$$

where $\hat{\mu}_{Aj}(\tau)$ and $\hat{\mu}_{Bj}(\tau)$ are estimators for $\mu_j(\tau)$, $j = 1, 2$. Let $\hat{\Psi}_{ji}$ be straightforward modification of $\hat{\Psi}_i$ given in Theorem 3.1 and define $\hat{\pi}_j = n_j/(n_1 + n_2)$. With the regularity conditions in Theorem 3.1 being satisfied for each group, under $H_0: \mu_1(\tau) = \mu_2(\tau)$, $\{n_1 n_2 / (n_1 + n_2)\}^{1/2} W_A$ and $\{n_1 n_2 / (n_1 + n_2)\}^{1/2} W_B$ converge in distribution to a zero-mean normal random variable whose variance can be consistently estimated by

$$\hat{\pi}_2 \sum_{i=1}^{n_1} \hat{\Psi}_{1i}(\tau)^2 / n_1 + \hat{\pi}_1 \sum_{i=1}^{n_2} \hat{\Psi}_{2i}(\tau)^2 / n_2. \text{ When setting } Y_{ji}(\cdot) = 1 \text{ for } i = 1, \dots, n_j \text{ and } j = 1, 2,$$

W_A reduces to the weighted Kaplan–Meier statistic $WKM = \int_0^\tau w(t) \{\hat{S}_{D1}(t) - \hat{S}_{D2}(t)\} dt$ (Pepe and Fleming 1989).

The value of τ is determined based on scientific interest and is usually left to domain experts. A larger value of τ often leads to a greater variability in the estimation of $\mu(\tau)$ (as suggested by Table 1 in simulation studies) due to limited observation of marker process near τ . On the other hand, the summary measure $\mu(\tau)$ with a small value of τ does not fully use available data and does not properly characterize the long term treatment effect. When a relatively large value of τ is chosen to incorporate more data information, we recommend using a decreasing weight function $w(t)$ to reduce the variance. Consider a randomized clinical trial where a total of n_1 and n_2 subjects were assigned to two study arms. Assume that patients can potentially survive beyond the end of the study period τ_C , that is, $\Pr(D > \tau_C) > 0$. It is known that, when setting $w(\cdot) = 1$, $\{n_1 n_2 / (n_1 + n_2)\}^{1/2} \{\hat{\mu}_1(\tau_C) - \hat{\mu}_2(\tau_C)\}$ is not asymptotically bounded in probability, where $\hat{\mu}_1(\tau_C)$ and $\hat{\mu}_2(\tau_C)$ are the estimated summary measures for the two study arms (Pepe and Fleming 1991). In this case, we can use a random weight function such as

$$\hat{w}(t) = \frac{(n_1 + n_2) \hat{S}_{C1}(t) \hat{S}_{C2}(t)}{n_1 \hat{S}_{C1}(t) + n_2 \hat{S}_{C2}(t)} \quad (3.5)$$

to downweight the difference in the later time period, where $\hat{S}_{C1}$ and $\hat{S}_{C2}$ are the Kaplan–Meier estimators of the survival function of the censoring time in the two arms. The corresponding test statistic is

$$\hat{W} = \int_0^\tau \hat{w}(t) \hat{E}\{Y_1(u)I(D_1 > u)du\} - \int_0^\tau \hat{w}(t) \hat{E}\{Y_2(u)I(D_2 > u)du\}, \text{ with } \tau = \min(\hat{\tau}_{C1}, \hat{\tau}_{C2}) \text{ and } \hat{\tau}_{Ci} = \sup\{u: \hat{S}_{Ci}(u) > 0\} \text{ for } i = 1, 2. \text{ This simple weight function ensures that the asymptotic variance of } \{n_1 n_2 / (n_1 + n_2)\}^{1/2} \hat{W} \text{ is finite.}$$

4. Nonparametric Estimation of $\mu(t)$ when Marker is Continuously Observed

In this section, we consider estimation of $\mu(t)$ when the longitudinal marker process $Y(\cdot)$ is completely observed before the terminal event or censoring. The observed data $\{X_i, \delta_i, I(X_i \leq t) Y_i(t) : 0 \leq t \leq \tau, i = 1, \dots, n\}$ are assumed to independent replicates of $\{X, \delta, I(X \leq t) Y(t) : 0 \leq t \leq \tau\}$. As in Section 3, the key step is to estimate the function $r(u) = E\{Y(u) / D(u)\}$. Under the independent censoring assumption, for $u \in [0, \tau]$, we propose to estimate $r(u)$ by the moment type estimator

$$\tilde{r}(u) = \frac{\sum_{i=1}^n Y_i(u) I(X_i \geq u)}{\sum_{i=1}^n I(X_i \geq u)}.$$

Thus a straightforward estimator of $\mu(t)$ is

$$\tilde{\mu}(t) = \int_0^t w(u) \hat{S}_D(u) \tilde{r}(u) du.$$

Note that the moment-type estimator $\tilde{r}(u)$ is a $\sqrt{n}$ -consistent estimator for $r(u)$, while the kernel-type estimator $\tilde{r}_h(u)$, in (3.2) has a $\sqrt{nh}$ convergence rate. Interestingly, $\tilde{\mu}(t)$ can be shown to be more efficient than $\tilde{\mu}_A(t)$ and $\tilde{\mu}_B(t)$. Theorem 4.1 states the asymptotic properties of $\tilde{\mu}(t)$, with proof given in supplementary materials.

Theorem 4.1

Under Assumptions (A1) and (A2) in the Appendix, the stochastic process

$n^{1/2}\{\tilde{\mu}(t) - \mu(t)\} (0 \leq t \leq \tau)$ has an asymptotically iid representation

$n^{1/2}\{\tilde{\mu}(t) - \mu(t)\} = n^{-1/2} \sum_{i=1}^n U_i(t) + o_p(1)$, where

$$U_i(t) = \int_0^t \frac{\mu(u) dM_i^D(u)}{S_X(u)} - \mu(t) \int_0^t \frac{dM_i^D(u)}{S_X(u)} + \int_0^t \frac{w(u) S_D(u) Y_i(u) I(X_i \geq u)}{S_X(u)} du \\ - \int_0^t \frac{w(u) S_D(u) E\{Y(u) I(X \geq u)\} I(X_i \geq u)}{S_X(u)^2} du.$$

Moreover, as $n \rightarrow \infty$, $\sqrt{n}\{\tilde{\mu}(t) - \mu(t)\} (0 \leq t \leq \tau)$ converges weakly to a zero mean Gaussian process with the variance-covariance function $E\{U_1(s)U_1(t)\}$. Moreover, $E\{U_1(t)\}^2 = E\{\Psi_1(t)\}^2$ for all $t \in [0, \tau]$.

Similar as the arguments in Section 3, the variance-covariance function $E\{U_1(s)U_1(t)\}$ can be estimated by $1/n \sum_{i=1}^n \hat{U}_i(s)\hat{U}_i(t)$, with $\hat{U}_i$ defined in the supplementary materials. For testing the null hypothesis $H_0: \mu_1(\tau) = \mu_2(\tau)$, with subscript j ($j = 1, 2$) indicating group number as in Section 3, we consider a straightforward test statistic $\tilde{W} = \tilde{\mu}_1(\tau) - \tilde{\mu}_2(\tau)$. Let $\hat{U}_{ij}$ be straightforward modification of $\hat{U}_i$, $j = 1, 2$. With the regularity conditions in Theorem 4.1 being satisfied for each group, under H_0 , $\{n_1 n_2 / (n_1 + n_2)\}^{1/2} \tilde{W}$ converge weakly to a normal random variable, whose variance can be consistently estimated by

$$\hat{\pi}_2 \sum_{i=1}^{n_1} \hat{U}_{1i}(\tau)^2 / n_1 + \hat{\pi}_1 \sum_{i=1}^{n_2} \hat{U}_{2i}(\tau)^2 / n_2.$$

An important application of the proposed methods is benefit-risk assessment that combines information from a multiple event process and a terminal event (see Example 2.2 in Section 2). Let $O(\cdot)$ denote the multiple event counting process that increase by one when a non-terminal event occurs. For ease of discussion, assume that a smaller value of $O(\cdot)$ at any time point is preferred. To perform benefit-risk assessment based on $\{X_i, I(X_i \leq t) O(t) : 0 \leq t \leq \tau\}$, we set $Y(t)$ to be a function of $O(t)$, say, $Y(t) = f\{O(t)\}$, where f is a prespecified nonincreasing function with $f(\cdot) \geq 0$. In this case, $Y(t)$ can be viewed as a score that characterizes patient's disease burden and health condition, and a larger value of $Y(t)$ is

desired. Without loss of generality, we set $w(\cdot) = 1$ since the weight function can be absorbed into f . In practice, the function f can be determined by the investigators. We consider two choices of f for illustration. A simple approach is to define a truncated reverse counting process with $f(x) = (K - x) I(K - x) + 1$, where K is a prespecified integer. In this way, only the first K non-terminal events are of interest and $Y(\cdot)$ stays 1 after the K th event until the terminal event occurs. Another approach is to define $f(x) = a^x$, where $0 < a < 1$. Then each subject starts with a score of 1, and the occurrence of a nonterminal event at time t discounts a patient's score $Y(t)$ by a factor of a . In contrast with the truncated reverse counting process approach, all the non-terminal events are of interest. We recommend the use of second approach when the number of event of interest that can be potentially observed is not fixed.

5. Simulation Studies

We conducted simulation studies to examine the finite-sample performance of the proposed methodology. The results for the estimators in Section 3 are presented in Section 5.1, where we assume the potential observation time process $N^*(\cdot)$ is independent with $\{D, C, Y(\cdot)\}$. In practice, the potential observation time process $N^*(\cdot)$ are sometimes correlated with $\{Y(\cdot), D\}$, and the methods in Section 3 may fail to provide consistent estimates for $\mu(\tau)$. To circumvent this problem, we consider a new estimating approach in Section 5.2, based on a regression model that allows the dependence of observation time process, marker process, and terminal event time through covariates. The simulation results for the new approach are presented in Section 5.2. The simulation studies for two-sample tests of Section 3 and 4 with different weight functions are presented in supplementary materials.

5.1. Simulation for Independent Observation Times

When $Y(\cdot)$ is intermittently observed, we examine the performance of $\hat{\mu}_A(t)$, $\hat{\mu}_B(t)$ for one-sample estimation. In the following simulations, the association between D and $Y(\cdot)$ is induced by a shared subject-specific random effect V , where V is generated from a normal distribution with mean 0 and variance σ_1^2 . Specifically, given V , the terminal event time D is generated from exponential distribution with rate parameter $\lambda = a_0 + (V + k\sigma_1)^2$. Moreover, the longitudinal marker process is generated from $Y(t) = g(t) + V + \varepsilon(t)$; where the error term $\varepsilon(t)$ is a mean zero Gaussian process with independent increments and a time-invariant variance σ_2^2 . Straightforward algebra gives $E\{Y(t)|D \geq t\} = g(t) - 2k\sigma_1^3 t / (1 + 2\sigma_1^2 t)$ and $\Pr(D \geq t) = (1 + \sigma_1^2 t)^{-1/2} \exp(-a_0 t - k^2 \sigma_1^2 t / (1 + 2\sigma_1^2 t))$. Note that when $k = 0$, survivors' expected marker value $E\{Y(t) | D \geq t\}$ is $g(t)$, which is the same as $E\{Y(t)\}$; moreover, the difference between $E\{Y(t) | D \geq t\}$ and $E\{Y(t)\}$ becomes larger as $|k|$ increases. The model implies that subjects with V close to $-k\sigma_1$ have smaller rate parameter for the terminal event, and tend to have longer survival time.

In our simulations for one-sample estimation, we set $k = 1$, $a_0 = 0.1$, $\sigma_1 = 0.5$, $\sigma_2 = 0.1$, $w(t) = 1$, $g(t) = t + 1$. The censoring time is generated from the uniform distribution on $[0, 5]$. The observation times are generated from $I(X \leq t) dN^*(t)$ and $N^*(t)$ is a Poisson process with

constant rate $\lambda^*(t) = 5$. We examine the performance of $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ when using the Epanechnikov kernel with bandwidth $h = n^{-1/3}$, $n^{-2/5}$, as the regularity condition (A5) given in the Appendix is $h \propto n^{-\nu}$, $1/4 < \nu < 1/2$. For the problem of bandwidth selection, the real need is to get bandwidths that satisfy (A5), and a precise target of an optimal bandwidth is often unnecessary. The averaged square error (Härdle et al. 2004) is a commonly used criteria that measures how close the estimate $\hat{r}_h$ is to the true curve r , which can be defined

as $n^{-1} \sum_{i=1}^n \int_0^t \{r(u) - \hat{r}_h(u)\}^2 dN_i(u)$ in our case. Note that minimizing

$CV(h) = n^{-1} \sum_{i=1}^n \int_0^t \{Y_i(u) - \hat{r}_{h,-i}(u)\}^2 dN_i(u)$, where $\hat{r}_{h,-i}(u)$ is the estimate for $r(u)$

leaving out the i th observation, is on average equivalent to minimizing

$n^{-1} \sum_{i=1}^n \int_0^t \{r(u) - \hat{r}_h(u)\}^2 dN_i(u)$. The leave-one-out cross-validation procedure yields a

bandwidth $\propto n^{-1/(2s+1)}$, where s is the order of kernel function used in cross-validation. By using a first order kernel, for example, $K_{1h}(u) = I(0 \leq u \leq h)/h$, to calculate $CV(h)$, the optimal bandwidth $\hat{h} = \hat{c}n^{-1/3}$ satisfies the regularity condition. We then use a bounded symmetric kernel function, say, the Epanechnikov kernel, with the selected bandwidth $\hat{h}$ for nonparametric estimation. The procedure shares similar ideas to those mentioned in Newey, Hsieh, and Robins (2004) and Maity, Ma, and Carroll (2007). Table 1 presents the summary statistics for $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$ based on 2000 replications. From the simulation studies, if the bandwidth satisfies the technical conditions (A5), the performance of the proposed estimator is almost not affected by using different bandwidths ($h = n^{-1/3}$, $h = n^{-2/5}$ and data-adaptive bandwidth) when sample sizes are relatively large. Moreover, the simulation results under scenarios with $k = 0$ and 2 are presented in Supplemental Materials. Our proposed procedure performs well in finite-sample studies.

5.2. Simulation for Dependent Observation Times

We now relax the assumption and consider the case where the association between the marker process $Y(\cdot)$, time to the terminal event D and the observation time process $N^*(\cdot)$ is characterized by a vector of covariates Z . We assume that $N^*(\cdot)$ and C are independent with $\{Y(\cdot), D\}$ conditional on Z . Let $h(t|Z)$ be the hazard function of the terminal event time D given Z . We consider the following regression models

$$h(t|Z) = \exp(\eta'Z)h_0(t), \quad (5.1)$$

$$E\{Y(t)|D \geq t, Z\} = \alpha_0(t) + \beta'Z, \quad (5.2)$$

$$E\{dN^*(t)|Z\} = \exp(\gamma'Z)\lambda_0(t)dt, \quad (5.3)$$

where $h_0(t)$, $\alpha_0(t)$, and $\lambda_0(t)$ are unspecified baseline functions. Specifically, Model (5.1) is the Cox proportional hazards model; Model (5.2) characterizes the mean function of marker process among survivors while leaving the distributional form unspecified, and the effects of covariates are characterized by a linear relationship; Model (5.3) formulates a proportional rate model for the observation time process, where the baseline rate function $\lambda_0(t)$ is left unspecified. Under Models (5.1) and (5.2), the summary measure in the regression setting can be defined as $\mu(\tau|Z) = \int_0^\tau \{\alpha_0(u) + \beta'Z\} \exp\{-H_0(u)\exp(\eta'Z)\} du$, with $H_0(t) = \int_0^t h_0(u) du$.

Thus $\mu(\tau)$ can be expressed as $\mu(\tau) = \int_{-\infty}^\infty \mu(\tau|z) F_Z(dz)$, where $F_Z(z)$ is the cumulative distribution function of Z . When there are no covariate effects, that is, $\eta = \beta = \gamma = 0$, the model reduces to the original nonparametric model in Section 3.

Note that the marker process and observation times are correlated through Z , we propose to estimate β and γ in Models (5.2) and (5.3) by applying the methods in Lin and Ying (2001). Specifically, Let $(\hat{\beta}, \hat{\gamma})$ be the solution to the estimating equations

$$\sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(u, \gamma)\} dN_i(u) = 0, \quad \sum_{i=1}^n \int_0^\infty \{Z_i - \bar{Z}(u, \gamma)\} \{Y_i(u) - \beta'Z_i\} dN_i(u) = 0,$$

where $\bar{Z}(u, \gamma) = \frac{\sum_{i=1}^n Z_i I(X_i \geq t) \exp(\gamma'Z_i)}{\sum_{i=1}^n I(X_i \geq t) \exp(\gamma'Z_i)}$. We propose to estimate $\alpha_0(t)$ by a kernel type estimator

$$\hat{\alpha}_0(t) = \frac{\sum_{i=1}^n \int_0^\tau K_h(t-u) \hat{\varepsilon}_i(u) dN_i(u)}{\sum_{i=1}^n \int_0^\tau K_h(t-u) dN_i(u)}, \quad t \in [h, \tau-h],$$

with $\hat{\varepsilon}_i(u) = Y_i(u) - \hat{\beta}'Z_i$. Moreover, we set $\hat{\alpha}_0(t) = \hat{\alpha}_0(h)$ for $t \in [0, h]$ and set $\hat{\alpha}_0(t) = \hat{\alpha}_0(\tau-h)$ for $t \in [\tau-h, \tau]$. Next, let $\hat{\eta}$ be the partial likelihood estimator for η and let $\hat{H}_0(t)$ be the Breslow-type estimator for $H_0(t)$. Then a consistent estimator of $\mu(\tau)$ can be given by

$$\hat{\mu}_R(\tau) = \frac{1}{n} \sum_{i=1}^n \int_0^\tau \{\hat{\alpha}_0(u) + \hat{\beta}'Z_i\} \exp\{-\hat{H}_0(u)\exp(\hat{\eta}'Z_i)\} du.$$

Simulation studies are conducted to evaluate the performance of $\hat{\mu}_R(\tau)$. Specifically, we set $\beta = \gamma = 1$, $\gamma = 2$, $\alpha_0(t) = t + 1$ and $h_0(t) = 2t$, and generate $Y(t) = \alpha_0(t) + \beta Z + \varepsilon(t)$, where Z follows a uniform distribution on $[0, 1]$, and $\varepsilon(t)$ is a mean zero Gaussian process with independent increments and a time-invariant variance 0.04. The baseline rate function of N^* is chosen to be $\lambda_0(t) = 1$ or $\lambda_0(t) = 0.5 + 0.5e^{-t}$. The simulation results presented in Table 2. The proposed estimator works well with finite sample sizes.

6. Data Examples

6.1. Primary Biliary Cirrhosis Clinical Trial

We illustrate the proposed methods through an application to a primary biliary cirrhosis (PBC) clinical trial (Murtaugh et al. 1994). In this trial, patients were enrolled for evaluating the use of D-penicillamine for to treat PBC. As discussed in Murtaugh et al. (1994), no significant benefit of D-penicillamine was found. Of the 276 female patients, 137 were randomized to receive D-penicillamine, and 139 was randomized to the placebo group. Patient accrual took place from January 1974 through May 1984, and follow-up was extended to April 1988, by which time 114 of the female patients had died. Aspartate transaminase (AST), with elevated levels indicating liver damage or disease, was recorded repeatedly at intermittent time points. Measurements of AST were scheduled to be collected at a annually basis, however, the real observation times were unbalanced due to mistimed measurements, skipped visits and dropouts. We investigate the effect of treatment based on AST and time to death within five years.

We define Y to be a function of AST value at each time point t . Because the normal range of AST is $0 \sim 40$ IU/mL and most PBC patients have AST less than 400 IU/mL, we set $Y(t) = 1$ for $AST(t) \leq 40$, and set $Y(t) = 0$ for $AST(t) > 400$. When $40 < AST < 400$, we set Y to be a decreasing linear function of AST from 1 to 0, that is, $Y(t) = 1 - (AST(t) - 40)/360$. By taking $w(\cdot) = 1$ and applying methods in Section 3, the mean of the cumulative weighted marker process at $\tau = 5$ years is 3.606(SE, 0.124) for the D-penicillamine group and is 3.137(SE, 0.125) for the placebo group (SE stands for standard error). Figure 1 displays the estimated cumulative mean function $\mu(t)$, survival function and survivors' mean function in both groups within 5 years. The plots suggest that the difference in $\mu(\tau)$ is mainly explained by the larger survivor's mean function and prolonged survival of D-penicillamine group within five years. For two-sample test, using weight function $w(\cdot) = 1$ yields p -value 8.1×10^{-3} . Moreover, we investigate the immediate and long-term impacts of treatment by using different weight functions $w_1(u) = a^u$ and $w_2(u) = a^{\tau-u}$ ($0 < a < 1$). The plots of p -value of two-sample test with values of a ranging from 0 to 1 are presented in Figure 2. The left panel shows the p -value when the immediate outcome is assigned more weight; while the right panel shows the p -value when the long-term outcome is assigned more weight. Specifically, using weight function $w(u) = 0.95^u$ and $w(u) = 0.95^{5-u}$ yield p -value $8. \times 10^{-3}$ and 8.0×10^{-3} . The conclusion that D-penicillamine group performs better remains the same with different weight functions. Our analysis suggests that, at 5th year, although D-penicillamine does not provide a significant benefit on survivorship, it helps maintain a lower level of AST and a higher summary measure.

The above nonparametric estimation relies on the assumption that the potential sampling times are independent with time to death and AST measurements, which implies that skipped visits carry no information about patients' underlying medical condition, hence the visit times are uncorrelated with time to the end of follow-up (due to either death or dropout). In other words, the rate function $\lambda^*(t)$ of the potential observation time counting process $N^*(\cdot)$ is expected to be independent of $X = \min(D, C)$ under the independent sampling times assumption. We apply the two-step hypothesis testing procedure developed

in Wang and Huang (2014) to test the independence of the rate function $\lambda^*(t)$ on X . Independence of sampling times is rejected if the null hypothesis of rate independence is rejected. The testing procedure is performed by testing the shape independence of $\lambda^*(t)$ on X , followed by the test of size independence of $\lambda^*(t)$ on X . In all analyses, a significance level of 0.025 is assigned to each of the size- and shape-independence test, so that the overall Type I error rate for the rate independence of $\lambda^*(t)$ on X is controlled at 0.05. The shape- and size-independence tests yield p-values of 0.07 and 0.65 for the treatment group, and 0.60 and 0.58 for the control group. Hence the null hypothesis that λ^* is rate-independent of X is not rejected, and we do not reject the independent sampling time assumption.

In the case where the independence assumption of the sampling times may not hold, we also apply the methods based on regression models in Section 5.2 as a reference. The covariates Z are age, edema status, and biomarkers on the log scale including bilirubin, prothrombin time and albumin. The summary measure is 3.600 (SE, 0.114) for the treatment group and 3.213 (SE, 0.127) for the control group, which are close to the above nonparametric estimates. For two-sample testing, using different weight functions $w(u) = 1, 0.95^u, 0.95^{5-u}$ yields p-values 0.02, 0.02 and 0.03, respectively. The conclusions are consistent with the results using methods in Section 3.

6.2. CPCRA ddI/ddC Clinical Trial

We illustrate the proposed methods by analyzing data from a clinical trial conducted by Terry Bein Community Programs for Clinical Research on AIDS, a federally funded national network of community-based research groups. The study compared didanosine (ddI) and zalcitabine (ddC) as treatments for HIV-infected patients who were intolerant or had failed treatment with zidovudine. The trial randomized 230 patients to receive ddI treatment and 237 to receive ddC. The primary endpoint is time to disease progression or death. The secondary endpoints include changes in the Karnofsky performance score and opportunistic infections, where a reduction in the Karnofsky score and the occurrence of opportunistic disease indicate a deterioration in health. Both survival time and QoL are regarded as important indexes for treatment success. The analysis in Abrams et al. (1994) suggested that ddC treatment may have provided a survival advantage over ddI treatment, with borderline significance based on a proportional hazards model. We investigated the treatment effects on the cumulative weighted marker process for a more comprehensive assessment of the benefits and risks of the treatments. In our analysis, death is the terminal event of interest, and Karnofsky score and incidence of opportunistic infections are used as measures for QoL. The analysis with Karnofsky score illustrates the proposed methodology in the case where the longitudinal marker is intermittently observed, while the analysis with incidence of opportunistic infections illustrates the situation where the longitudinal marker is completely observed throughout the follow-up period.

In the first set of analysis, we divide the Karnofsky score by 100 and transform it into a 0 to 1 scale. By setting $w(\cdot) = 1$, the mean of the cumulative weighted marker process at $\tau = 1.37$ years (500 days) is 0.871 (SE, 0.021) for the ddI group and is 0.901 (SE, 0.021) for the ddC group. The two-sample test with $w(\cdot) = 1$ yields p -value 0.44. The plots of estimated cumulative mean function $\mu(t)$, survival function and mean Karnofsky score of survivors in

the ddI and ddC treatment groups are presented in supplementary materials. The plots show that ddC performs better in terms of survival but worse in terms of survivors' physical QoL, and the estimated summary measures for the two treatments are very close. Moreover, the plots of p-value with different weight functions as in Section 6.1 are also presented in supplementary materials.

In the second set of analysis, we consider benefit-risk assessment based on opportunistic infections and death. A total of 363 confirmed or probable opportunistic diseases indicating disease progression (Neaton et al. 1994) were reported. The number of opportunistic infections per subject ranges from 0 to 5, with median 1 and mean 0.78. Denote by $\alpha(u)$ the total number opportunistic infections occurred at or before time u , and set $Y(\cdot) = 0.8^{\alpha(\cdot)}$. Taking $w(\cdot) = 1$, the occurrence of opportunistic infection at time t discounts a patient's score $Y(t)$ by 0.8. Then, the estimated summary measure is 0.998 for the ddI group and 1.028 for the ddC group. The p -value derived from the proposed two-sample test is 0.27. Our analysis again suggests that ddC outperforms ddI in terms of the proposed summary measure on survival and opportunistic disease, although the advantage is not statistically significant.

7. Discussion

In this article, we consider benefit-risk assessment based on longitudinal marker measurements and time to event data. The proposed method is especially useful when conflict results about the treatment effects are reported for the two outcomes. Our estimation and testing procedures are more robust than the existing methods, such as Hwang, Tsauo, and Wang (1996), in the sense that the statistical procedures can be derived from one single dataset. Statistical inference properties are established for point estimate and hypothesis testing, hence the proposed methodology is expected to be attractive for practitioners to facilitate accurate decision-making.

The proposed methodologies have a wide range of application in biomedical and public health research. Besides the examples discussed in Section 2, the longitudinal measure $Y(\cdot)$ can also be a function of an surrogate biomarker for the survival outcome of interest; for example, CD4 cell count has been used as a surrogate for progression to AIDS or death in many AIDS studies. In the case where the follow-up duration is not long enough to accumulate adequate number of events for meaningful analysis, the clinical study may have insufficient power to detect treatment effects on the survival outcome. Compared with the conventional survival analysis, the proposed methods use additional information from the surrogate marker and possess the potential to increase power in detecting real treatment effects. Finally, instead of using a single marker process, a benefit-risk summary measure integrating multiple marker processes and time to event is under investigation.

When longitudinal measurements are observed or collected at only prespecified or regularly spaced sampling times, an ad-hoc approach is to carry forward the observed biomarker values. Specifically, suppose observation time points by design are $0 = t_0 < t_1 < \dots < t_K$, and all the marker measurements are collected only at the designed time points during the follow

up period. We define $Y(t) = Y(t_k)$ for $t \in [t_k, \min(D, t_{k+1}))$ when $t_k < D$, and the nonparametric methods in Section 4 can be used to estimate $\mu(\tau)$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors thank Dr. Lee-Jen Wei at Harvard University for sharing ideas on benefit–risk assessment research. The work in this article was partially motivated by a presentation he gave at Johns Hopkins University.

Funding

The research was supported by the National Institute of Health grants R01HL122212 and R01CA193888.

Appendix

Assumptions (A1)–(A5) are the regularity conditions:

- (A1) The censoring time C_i is independent of $\{D_i, N_i^*(\cdot), Y_i(\cdot)\}$ and $\Pr(X_i = \tau) > 0$.
- (A2) The marker process $Y_i(t)$ is bounded and nonnegative. The nonnegative weight function $w(t)$ is bounded and of bounded variation.
- (A3) The counting process $N_i^*(\cdot)$ is independent of $\{D_i, C_i, Y_i(\cdot)\}$. The observation time process $I(X_i \geq t)N_i^*(t)$ is bounded and the second derivative of its rate function $\lambda(t)$ is bounded. Moreover, $\lambda(t) > 0$ on $[0, \tau]$.
- (A4) Define $\xi(t)$ such that $\xi(t)dt = E[Y(t)I(X \geq t)dN^*(t)]$, the second derivative of $\xi(t)$ is bounded on $[0, \tau]$.
- (A5) $K(\cdot)$ is a symmetric kernel function with bounded support and bounded variation, and $h \propto n^{-\nu}$, $1/4 < \nu < 1/2$.

References

- Abrams DI, Goldman AI, Launer C, Korvick JA, Neaton JD, Crane LR, Grodesky M, Wakefield S, Muth K, Kornegay S, et al. A Comparative Trial of Didanosine or Zalcitabine After Treatment with Zidovudine in Patients With Human Immunodeficiency Virus Infection. *New England Journal of Medicine*. 1994; 330:657–662. [PubMed: 7906384]
- Claggett B, Uno H, Tian L, Wei LJ. A Reverse Counting Process for Analyzing Survival Data With Multiple Event Times. presented at the 2014 Joint Statistical Meetings. 2014
- Gelber R, Gelman R, Goldhirsch A. A Quality-of-Life-Oriented Endpoint for Comparing Therapies. *Biometrics*. 1989:781–795. [PubMed: 2790121]
- Glasziou PP, Cole BF, Gelber RD, Hilden J, Simes RJ. Quality Adjusted Survival Analysis With Repeated Quality of Life Measures. *Statistics in Medicine*. 1998; 17:1215–1229. [PubMed: 9670411]
- Glasziou PP, Simes RJ, Gelber RD. Quality Adjusted Survival Analysis. *Statistics in Medicine*. 1990; 9:1259–1276. [PubMed: 2277877]
- Gravelle H, Smith D. Discounting for Health Effects in Cost-Benefit and Cost-Effectiveness Analysis. *Health Economics*. 2001; 10:587–599. [PubMed: 11747043]

- Härdle, W., Sperlich, S., Werwatz, A., Müller, M. Nonparametric and Semiparametric Models. New York: Springer-Verlag; 2004.
- Henderson R, Diggle P, Dobson A. Joint Modelling of Longitudinal Measurements and Event Time Data. *Biostatistics*. 2000; 1:465–480. [PubMed: 12933568]
- Hogan JW, Laird NM. Mixture Models for the Joint Distribution of Repeated Measures and Event Times. *Statistics in Medicine*. 1997; 16:239–257. [PubMed: 9004395]
- Huang Y, Louis TA. Expressing Estimators of Expected Quality Adjusted Survival as Functions of Nelson–Aalen Estimators. *Lifetime Data Analysis*. 1999; 5:199–212. [PubMed: 10518370]
- Hwang JS, Tsao JY, Wang JD. Estimation of Expected Quality Adjusted Survival by Cross-Sectional Survey. *Statistics in Medicine*. 1996; 15:93–102. [PubMed: 8614748]
- Irwin J. The Standard Error of an Estimate of Expectational Life. *Journal of Hygiene*. 1949; 47:188–189. [PubMed: 15406758]
- Lin DY, Ying Z. Semiparametric and nonparametric regression analysis of longitudinal data. *Journal of the American Statistical Association*. 2001; 96:103–126.
- Maity A, Ma Y, Carroll RJ. Efficient Estimation of Population-Level Summaries in General Semiparametric Regression Models. *Journal of the American Statistical Association*. 2007; 102:123–139.
- Murray S, Cole B. Variance and Sample Size Calculations in Quality-of-Life-Adjusted Survival Analysis (Q-TWiST). *Biometrics*. 2000; 56:173–182. [PubMed: 10783793]
- Murtaugh PA, Dickson ER, Dam GMV, Malinchoc M, Grambsch PM, Langworthy AL, Gips CH. Primary Biliary Cirrhosis: Prediction of Short-Term Survival Based on Repeated Patient Visits. *Hepatology*. 1994; 20:126–134. [PubMed: 8020881]
- Neaton JD, Wentworth DN, Rhame F, Hogan C, Abrams DI, Deyton L. Considerations in Choice of a Clinical Endpoint for AIDS Clinical Trials. *Statistics in Medicine*. 1994; 13:2107–2125. [PubMed: 7846414]
- Newey WK, Hsieh F, Robins JM. Twicing Kernels and a Small Bias Property of Semiparametric Estimators. *Econometrica*. 2004; 72:947–962.
- Pepe MS, Fleming TR. Weighted Kaplan–Meier Statistics: A Class of Distance Tests for Censored Survival Data. *Biometrics*. 1989:497–507. [PubMed: 2765634]
- Pepe MS, Fleming TR. Weighted Kaplan–Meier Statistics: Large Sample and Optimality Considerations. *Journal of the Royal Statistical Society, Series B*. 1991; 53:341–352.
- Robberstad B. QALYs vs DALYs vs LYs Gained: What are the Differences, and What Difference Do They Make for Health Care Priority Setting? *Norsk Epidemiologi*. 2005; 15:183–191.
- Shen LZ, Pulkstenis E, Hoseyni M. Estimation of Mean Quality Adjusted Survival Time. *Statistics in Medicine*. 1999; 18:1541–1554. [PubMed: 10398291]
- Sun YQ, Wu HL. AUC-Based Tests for Nonparametric Functions with Longitudinal Data. *Statistica Sinica*. 2003; 13:593–612.
- The Beta-Blocker Evaluation of Survival Trial Investigators. A Trial of the Beta-Blocker Bucindolol in Patients With Advanced Chronic Heart Failure. *The New England Journal of Medicine*. 2001; 344:1659. [PubMed: 11386264]
- Tsiatis AA, Deguttola V, Wulfsohn MS. Modeling the Relationship of Survival to Longitudinal Data Measured With Error. Applications to Survival and CD4 Counts in Patients With AIDS. *Journal of the American Statistical Association*. 1995; 90:27–37.
- Viscusi W. Discounting Health Effects for Medical Decisions. *Valuing Health Care*. 1995; 7
- Walker D, Kumaranayake L. Allowing for Differential Timing in Cost Analyses: Discounting and Annualization. *Health Policy and Planning*. 2002; 17:112–118. [PubMed: 11861593]
- Wang MC, Chiang CT. Nonparametric Methods for Recurrent Event Data With Informative and Non-Informative Censorings. *Statistics in Medicine*. 2002; 21:445–456. [PubMed: 11813230]
- Wang MC, Huang CY. Statistical Inference Methods for Recurrent Event Processes With Shape and Size Parameters. *Biometrika*. 2014; 101:553–566. [PubMed: 26412863]
- Wang Y, Taylor JMG. Jointly Modeling Longitudinal and Event Time Data With Application to Acquired Immunodeficiency Syndrome. *Journal of the American Statistical Association*. 2001; 96:895–905.

- Weinstein MC, Stason WB. Foundations of Cost-Effectiveness Analysis For Health and Medical Practices. The New England Journal of Medicine. 1977; 296:716–721. [PubMed: 402576]
- Wu MC, Carroll RJ. Estimation and Comparison of Changes in the Presence of Informative Right Censoring by Modeling the Censoring Process. Biometrics. 1988; 44:175–188.
- Xu J, Zeger SL. Joint Analysis of Longitudinal Data Comprising Repeated Measures and Times to Events. Journal of the Royal Statistical Society, Series C. 2001; 50:375–387.
- Zhao H, Tsiatis AA. Efficient Estimation of the Distribution of Quality-Adjusted Survival Time. Biometrics. 1999; 55:1101–1107. [PubMed: 11315054]

Figure 1.

Estimated cumulative mean functions $\mu(t)$ using AST score and time to death (left), survival functions (middle) and mean AST score of survivors $E\{Y(t) / D = t$ (right) for D-penicillamine (solid line) and placebo (dashed line) groups.

Figure 2.
The p -value of two-sample test with $w_1(u) = a^u$ (left) and $w_2(u) = a^{\tau-u}$ for $0 < a < 1$.

Table 1Simulation summary statistics for $\hat{\mu}_A(t)$ and $\hat{\mu}_B(t)$.

$\mu(t)$	$\hat{\mu}_A(t)$			$\hat{\mu}_B(t)$		
	Bias	SE	CP	Bias	SE	SEE
$n = 100, h = n^{-1/3}$						
t = 1	1.076	0.015	0.062	0.936	0.015	0.062
t = 2	2.274	0.009	0.161	0.936	0.013	0.161
t = 3	3.540	0.001	0.308	0.934	0.013	0.308
$n = 100, h = n^{-2/5}$						
t = 1	1.076	0.007	0.062	0.938	0.008	0.062
t = 2	2.274	0.001	0.161	0.936	0.007	0.161
t = 3	3.540	0.012	0.309	0.932	0.009	0.308
$n = 100, \text{data-adaptive bandwidth}$						
t = 1	1.076	0.003	0.062	0.944	0.006	0.061
t = 2	2.274	0.006	0.160	0.938	0.005	0.161
t = 3	3.540	0.020	0.307	0.934	0.007	0.308
$n = 200, h = n^{-1/3}$						
t = 1	1.076	0.009	0.043	0.943	0.009	0.042
t = 2	2.274	0.005	0.111	0.950	0.008	0.110
t = 3	3.540	0.001	0.215	0.946	0.009	0.215
$n = 200, h = n^{-2/5}$						
t = 1	1.076	0.004	0.043	0.946	0.005	0.042
t = 2	2.274	0.001	0.110	0.951	0.004	0.110
t = 3	3.54	0.007	0.215	0.946	0.006	0.215
$n = 200, \text{data-adaptive bandwidth}$						
t = 1	1.076	0.002	0.043	0.949	0.004	0.042
t = 2	2.274	0.004	0.110	0.949	0.003	0.110
t = 3	3.540	0.011	0.214	0.947	0.006	0.215

Note: Bias is the empirical bias; SE is the empirical standard error; SEE is the empirical standard error; and the two estimators have the same SEE; CP is the empirical coverage probability of the 95% confidence interval.

Table 2

Simulation summary statistics for $\hat{\mu}_R(\tau)$.

	$h = n^{-1/3}$			$h = n^{-2/5}$				
	Bias	SE	SEE	CP	Bias	SE	SEE	CP
$\lambda_0(t) = 1, n = 100$								
$t = 1$	0.017	0.062	0.063	0.951	0.011	0.062	0.063	0.953
$t = 2$	0.019	0.187	0.189	0.949	0.014	0.187	0.189	0.952
$t = 3$	0.035	0.300	0.306	0.951	0.031	0.301	0.306	0.952
$\lambda_0(t) = 0.5 + 0.5e^{-t}, n = 100$								
$t = 1$	0.015	0.063	0.063	0.949	0.009	0.063	0.063	0.948
$t = 2$	0.020	0.191	0.190	0.943	0.016	0.191	0.190	0.944
$t = 3$	0.042	0.317	0.309	0.937	0.038	0.319	0.310	0.936
$\lambda_0(t) = 1, n = 200$								
$t = 1$	0.012	0.045	0.045	0.942	0.007	0.045	0.045	0.943
$t = 2$	0.018	0.138	0.134	0.945	0.014	0.138	0.134	0.942
$t = 3$	0.030	0.225	0.217	0.939	0.027	0.225	0.217	0.940
$\lambda_0(t) = 0.5 + 0.5e^{-t}, n = 200$								
$t = 1$	0.009	0.044	0.045	0.942	0.009	0.044	0.045	0.949
$t = 2$	0.014	0.133	0.134	0.955	0.013	0.132	0.134	0.953
$t = 3$	0.027	0.212	0.217	0.955	0.024	0.212	0.218	0.958

Note: Bias is the empirical bias; SE is the empirical standard error; SEE is the empirical mean of the bootstrapped standard error estimates; CP is the bootstrapped empirical coverage probability of the 95% confidence interval.