Ann Inst Stat Math (2012) 64:415–438
DOI 10.1007/s10463-010-0317-2

Hazard function estimation with cause-of-death data missing at random

Qihua Wang · Gregg E. Dinse · Chunling Liu

Received: 24 December 2008 / Revised: 8 January 2010 / Published online: 8 September 2010

Abstract Hazard function estimation is an important part of survival analysis. Interest often centers on estimating the hazard function associated with a particular cause of death. We propose three nonparametric kernel estimators for the hazard function, all of which are appropriate when death times are subject to random censorship and censoring indicators can be missing at random. Specifically, we present a regression surrogate estimator, an imputation estimator, and an inverse probability weighted estimator. All three estimators are uniformly strongly consistent and asymptotically normal. We derive asymptotic representations of the mean squared error and the mean integrated squared error for these estimators and we discuss a data-driven bandwidth selection method. A simulation study, conducted to assess finite sample behavior,

Q. Wang (✉)
Academy of Mathematics and Systems Science, Chinese Academy of Science,
Beijing 100080, China
e-mail: qhwang@amss.ac.cn

Q. Wang
Department of Mathematics and Statistics, Yunnan University, Kunming 650091, China

Q. Wang
Department of Statistics and Actuarial Science, The University of Hong Kong,
Pokfulam Road, Pokfulam, Hong Kong

G. E. Dinse
Biostatistics Branch, National Institute of Environmental Health Sciences,
P.O. Box 12233, Research Triangle Park, NC 27709, USA

C. Liu
Department of Applied Mathematics, Hong Kong Polytechnic University,
Hong Hum, Hong Kong

demonstrates that the proposed hazard estimators perform relatively well. We illustrate our methods with an analysis of some vascular disease data.

Keywords Imputation estimator · Inverse probability weighted estimator · Kernel estimator · Regression surrogate estimator

1 Introduction

A common feature of survival data is the presence of right censored observations. Censoring can occur, for example, if individuals withdraw from a study before dying or if a study ends before all subjects have died. Additionally, when multiple causes of death are operating, the time to death from one cause can be censored by a death from a different cause. For instance, in a clinical trial one might distinguish between deaths attributable to the disease of interest and deaths due to all other causes. Without loss of generality, we focus on a particular cause of death and treat all other causes as censoring mechanisms with respect to the death time of interest.

Let T and $C^{(0)}$ be random variables representing the time to death from the cause of interest and the time to usual (right) censoring, respectively. Let $T^{(1)}, T^{(2)}, \dots, T^{(r)}$ be the times to death from all other causes. In our problem, T may be censored by $C^{(0)}, T^{(1)}, \dots, T^{(r-1)}$ or $T^{(r)}$. Let $C = \min(C^{(0)}, T^{(1)}, \dots, T^{(r)})$, where C denotes the censoring random variable. We assume that T and C are independent and we observe $X = \min(T, C)$ and $\delta = I(T \le C)$, where $I(\cdot)$ is the indicator function. Let F, G, and L be the cumulative distribution functions for T, C, and X, respectively. Finally, let $\lambda(t) = lim_{\epsilon \to 0^+} P(t \le T < t + \epsilon | T \ge t)/\epsilon$ be the hazard function for T.

Censored survival time problems frequently are characterized in terms of hazard functions, and thus the estimation of $\lambda(t)$ has received much attention. Suppose the data consist of n independent and identically distributed pairs $\{(X_i, \delta_i) : i = 1, \dots, n\}$. One type of non-parametric hazard estimation is based on kernel smoothers of the form:

$$\lambda_n(t) = \frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \frac{\delta_i}{n - R_i + 1}\,, \tag{1}$$

where R_i is the rank of X_i, $K(\cdot)$ is a kernel function, and h_n is a sequence of bandwidths. Clearly, $\lambda_n(t)$ is a convolution of the kernel function and the nonparametric cumulative hazard estimator of Nelson (1972). This class of estimators has been investigated by several authors, including Blum and Susarla (1980), Tanner (1983), Ramlau-Hansen (1983), Tanner and Wong (1983), Regina and John (1985), Diehl and Stute (1988), and Wang (1999).

This paper addresses the problem in which cause of death is unknown for a subset of individuals, and thus some of the censoring indicators are missing. For example, van der Laan and McKeague (1998) describe epidemiological studies in which death certificates were missing for some people, mainly due to emigration or inconclusive hospital case notes and autopsy results. They point out that it can be impossible to determine whether death was due to the cause of interest in these cases. Missing causes of death also arise in carcinogenicity experiments. In some studies only a subset of animals are examined for tumors to cut costs; occasionally tissues autolyze or are cannibalized by cage mates before a necropsy can be performed; and pathologists are not always able to determine each tumor's role in causing death. For example, Kalbfleisch and Prentice (1980) provide data on mice that died from leukemia, other known (non-leukemia) causes, or unknown causes; and Dinse (1986) presents data on mice whose status of non-renal vascular disease at death was classified as absent, incidental, fatal, or unknown. This last data set is analyzed in Sect. 6.

The general problem of analyzing censored survival data with missing cause-of-death data (or missing censoring indicators) has received much attention. Dinse (1982) derived the nonparametric maximum likelihood estimator of the survival function in this situation; see, also, the estimators of Dinse (1986), Lo (1991), McKeague and Subramanian (1998), van der Laan and McKeague (1998), and Subramanian (2004, 2006). Other authors have considered hypothesis testing and regression modeling. Goetghebeur and Ryan (1990) derived a modified log-rank test to compare survival rates in two groups; Dewanji (1992) suggested an improvement to that approach; and Goetghebeur and Ryan (1995) extended their earlier results to the proportional hazards regression model. Tsiatis et al. (2002) used multiple imputation methods to evaluate treatment differences in survival. Recently, Gao and Tsaitis (2005) developed a semi-parametric procedure to estimate regression coefficients in a linear transformation competing risks model. Klein and Moeschberger (2003, Chapter 6) point out the importance of kernel estimation for hazard functions in the presence of censored data. In this paper, we concentrate on non-parametrically estimating the hazard function, $\lambda(t)$, by extending well-known kernel smoothing methods to allow for missing data.

Suppose that X is always observed, but the censoring indicator δ is missing for some subjects. Define a missingness indicator ξ which is 1 if δ is observed and is 0 otherwise. Therefore, we observe either $\{X, \delta, \xi = 1\}$ or $\{X, \xi = 0\}$. Throughout this paper, we assume that δ is missing at random (MAR), which implies that ξ and δ are conditionally independent given X: $P(\xi = 1|X, \delta) = P(\xi = 1|X)$. The MAR assumption is common in statistical analyses involving missing data and is reasonable in many practical situations; see, for example, Little and Rubin (1987, Chapter 1).

When some censoring indicators are missing, the hazard estimator in (1) cannot be applied directly. One simple solution is to use only the complete cases, $\{X, \delta, \xi = 1\}$, and to ignore all subjects with missing indicators, $\{X, \xi = 0\}$. However, the resulting complete case (CC) estimator is highly inefficient if there is a significant degree of missingness; see, e.g., van der Laan and McKeague (1998). Also, the CC estimator is consistent and unbiased only when the censoring indicators are missing completely at random (MCAR), which is a special case of MAR where ξ is independent of both X and δ: $P(\xi = 1|X, \delta) = P(\xi = 1)$; see, e.g., Tsiatis et al. (2002).

Imputation has become a popular method for handling missing data; see, for example, Rubin (1987), Lipsitz et al. (1998), Robins and Wang (2000), and Wang and Rao (2002). The popularity of this approach stems largely from the fact that once the missing values are imputed, standard techniques for analyzing complete data can be readily applied. The inverse probability weighted procedure is also widely used in missing data situations; see, for example, Robins and Rotnitzky (1992); Robins et al. (1994), and Zhao et al. (1996). These two approaches are usually applied to regression problems with missing responses or covariates, but here we adapt them to handle missing censoring indicators.

This paper develops three kernel estimators for the hazard function: a regression surrogate estimator, an imputation estimator, and an inverse probability weighted estimator. The regression surrogate estimator is based on a particular expression for $\lambda(t)$, and the imputation estimator and inverse probability weighted estimator are motivated by the regression surrogate estimator. All three estimators are of the form given in (1), except that some or all of the δ_i values are replaced by other quantities. The regression surrogate estimator replaces every δ_i, known or unknown, by an estimator of the conditional expectation of δ_i given X_i. The imputation estimator replaces only the unknown values of δ_i by estimators of their conditional expectations. The inverse probability weighted estimator replaces each unknown δ_i by its estimated conditional expectation and each known δ_i by a weighted sum of δ_i and its estimated conditional expectation. Existing non-parametric methods for estimating hazard functions assume complete censoring information, while non-parametric methods that allow missing censoring indicators focus on the simpler problem of estimating survival functions rather than hazard functions. The main contributions of this paper are the development of non-parametric approaches for estimating hazard functions with missing censoring indicators and the theoretical results derived for the proposed estimators.

The paper is organized as follows. Section 2 defines three nonparametric hazard estimators. Section 3 shows that these estimators are uniformly strongly consistent and asymptotically normal under the MAR assumption, and derives asymptotic representations for the mean squared error (MSE) and mean integrated squared error (MISE). Section 4 gives a data-driven bandwidth selection procedure. Section 5 reports simulation results for evaluating finite sample performance; Sect. 6 illustrates our methods by applying them to data from an animal experiment; and Sect. 7 provides a few concluding remarks. Finally, the main results are proved in the Appendices.

2 Estimation

The hazard function of interest, $\lambda(t)$, can be expressed as

$$\lambda(t) = \frac{f(t)}{1-F(t)} = \frac{[1-G(t)]f(t)}{1-L(t)} = \frac{1}{1-L(t)}\frac{\mathrm{d}L_1(t)}{\mathrm{d}t}\,, \tag{2}$$

where $f(t) = \mathrm{d}F(t)/\mathrm{d}t$ and $L_1(t) = P(X \le t, \delta = 1)$. As noted by Dikta (1998), we can write $L_1(t) = \int_0^t m(s)\mathrm{d}L(s)$, where $m(s) = P(\delta = 1|X = s)$ is the conditional expectation of the censoring indicator given the observation time, and thus (2) yields

$$\lambda(t) = \frac{m(t)}{1-L(t)}\frac{\mathrm{d}L(t)}{\mathrm{d}t}. \tag{3}$$

Given (3), we use kernel smoothing to define a regression surrogate estimator of $\lambda(t)$:

$$\widehat{\lambda}_{n,S}(t) = \frac{1}{h_n}\int K\left(\frac{t-s}{h_n}\right)\frac{m_n(s)\,\mathrm{d}L_n(s)}{1-L_n(s-)}\,, \tag{4}$$

where $m_n(s) = \sum_{i=1}^n \xi_i \delta_i W\left(\frac{s-X_i}{b_n}\right) / \sum_{i=1}^n \xi_i W\left(\frac{s-X_i}{b_n}\right)$, $L_n(s) = n^{-1}\sum_{i=1}^n I(X_i \leq s)$, and $s-$ is the time just before s. Here $m_n(s)$ is the Nadaraya–Watson kernel regression estimator of $m(s)$, where $W(\cdot)$ is a kernel function and b_n is a bandwidth sequence.

As $n[1 - L_n(X_i-)] = n - R_i + 1$, the estimator in (4) can be rewritten as

$$\widehat{\lambda}_{n,S}(t) = \frac{1}{h_n}\sum_{i=1}^n K\left(\frac{t - X_i}{h_n}\right)\frac{m_n(X_i)}{n - R_i + 1}. \tag{5}$$

If no censoring indicators are missing, the regression surrogate estimator in (5) reduces to a pre-smoothed Nelson–Aalen type estimator (Cao et al. 2005; Cao and Jácome 2004; Jacome et al. 2008). Similarly, the basic kernel estimator in (1), which is appropriate when none of the censoring indicators are missing, coincides with the regression surrogate estimator in (5) if every δ_i is replaced by $m_n(X_i)$, an estimator of the conditional expectation of δ_i given X_i. Intuitively, however, it seems reasonable to replace only the missing censoring indicators with estimators of their conditional expectations. If we use this logic and only replace δ_i by $m_n(X_i)$ in (1) if $\xi_i = 0$, we obtain our imputation estimator of $\lambda(t)$:

$$\widehat{\lambda}_{n,I}(t) = \frac{1}{h_n}\sum_{i=1}^n K\left(\frac{t - X_i}{h_n}\right)\frac{\xi_i\,\delta_i\; +\; (1 - \xi_i)\, m_n(X_i)}{n - R_i + 1}. \tag{6}$$

Finally, define $\pi(x) = P(\xi = 1|X = x)$, plus its Nadaraya–Watson kernel regression estimator $\pi_n(x) = \sum_{i=1}^n \xi_i\, \Omega\left(\frac{x-X_i}{\gamma_n}\right) / \sum_{i=1}^n \Omega\left(\frac{x-X_i}{\gamma_n}\right)$, where $\Omega(\cdot)$ is a kernel function and γ_n is a bandwidth sequence. Our inverse probability weighted estimator of $\lambda(t)$ is

$$\widehat{\lambda}_{n,W}(t) = \frac{1}{h_n}\sum_{i=1}^n K\left(\frac{t - X_i}{h_n}\right)\frac{[\xi_i/\pi_n(X_i)]\,\delta_i\; +\; [1 - \xi_i/\pi_n(X_i)]\, m_n(X_i)}{n - R_i + 1}. \tag{7}$$

In this case, δ_i in (1) is replaced by $m_n(X_i)$ if $\xi_i = 0$ and by a weighted average of δ_i and $m_n(X_i)$ if $\xi_i = 1$, where each weight is inversely proportional to $\pi_n(X_i)$, which is an estimator of the conditional expectation of ξ_i given X_i.

Just like other nonparametric kernel methods, our methods are robust to the choice of kernel functions. Some common kernel functions include the uniform, biweight, and Epanechnikov kernel functions. Bandwidth selection for h_n and robustness of our methods with respect to b_n and γ_n are discussed in subsequent sections.

All three of our hazard estimators use a non-parametric estimator for $m(x)$. One natural alternative is a semi-parametric approach that assumes a specific parametric model, say $m(x|\theta)$, where θ is a finite-dimensional parameter and $m(\cdot|\cdot)$ is a known function. Another alternative for handling missing data is the use of multiple imputation. Conditional on any observation time X_i for which the censoring indicator is missing ($\xi_i = 0$), generate ν independent Bernoulli random variables, say $\{\delta_{ij}^* : j = 1, \ldots, \nu\}$, each of which is 1 with probability $m_n(X_i)$. An individual imputation estimator, say $\lambda_{n,j}(t)$, can be obtained by replacing each missing δ_i with

δ^*_{ij} in (1) for $j = 1, \ldots, \nu$. As this is equivalent to replacing $m_n(X_i)$ with δ^*_{ij} in (6), we could define two other estimators by performing the same replacement in (5) and (7). In any of these three cases, the multiple imputation estimator is the average: $\nu^{-1}\sum_{j=1}^{\nu}\lambda_{n,j}(t)$. We plan to investigate these two approaches in a separate paper.

3 Asymptotic properties

This section discusses asymptotic properties of the estimators proposed in Sect. 2. Let $\widehat{\lambda}_n(t)$ denote any one of the estimators in (5)–(7). The following theorem, which is proved in Appendix A, establishes the uniform strong consistency of $\widehat{\lambda}_n(t)$.

Theorem 1 *Under the assumptions given in Appendix A, we have*

$$\sup_{0\le t\le \tau}\left|\widehat{\lambda}_n(t)-\lambda(t)\right| \xrightarrow{a.s.} 0\,,$$

where $0 < \tau < \tau_L$ *and* $\tau_L = \inf\{t : L(t) = 1\}$.

The next theorem, which is proved in Appendix B, establishes the asymptotic normality of $\widehat{\lambda}_n(t)$.

Theorem 2 *Under the assumptions given in Appendix B, we have*

$$\sqrt{nh_n}\left(\widehat{\lambda}_n(t)-\lambda(t)-(-1)^k\frac{\lambda^{(k)}(t)h_n^k}{k!}\int u^k K(u)\,du\right)\xrightarrow{\mathcal{L}} N\left(0,\sigma^2(t)\right)$$

for any fixed $0 < t < \tau_L$, *where* $\lambda^{(k)}(t)$ *is the kth derivative of* $\lambda(t)$.

The asymptotic variance, $\sigma^2(t)$, in the above theorem is

$$\sigma^2(t) = \frac{\lambda(t)}{1-L(t)}\int K^2(u)\,\mathrm{d}u + \frac{[1-\pi(t)]m(t)[1-m(t)]l(t)}{\pi(t)[1-L(t)]^2}\int K^2(u)\,\mathrm{d}u, \quad (8)$$

where $l(t) = \mathrm{d}L(t)/\mathrm{d}t$. A consistent estimator of $\sigma^2(t)$, say $\widehat{\sigma}^2_n(t)$, can be obtained by replacing $\lambda(t)$, $L(t)$, $\pi(t)$, $m(t)$ and $l(t)$ in (8) by estimators $\widehat{\lambda}_n(t)$, $L_n(t)$, $\pi_n(t)$, $m_n(t)$ and $l_n(t)$, where $l_n(t)$ is the kernel density estimator $l_n(t) = (n\,h_n)^{-1}\sum_{i=1}^{n}K\left(\frac{t-X_i}{h_n}\right)$. It is easy to see that the asymptotic variance reduces to that of the standard kernel hazard function estimator $\lambda_n(t)$ in (1) if there are no missing censoring indicators, in which case $\pi(t) = 1$.

The previous theorem implies that the asymptotically optimal bandwidth for a fixed value of t, which minimizes the asymptotic mean squared error, is

$$h_{opt,t} = \left(\frac{\sigma^2(t)}{[\lambda^{(k)}(t)/k!]^2\,[\int u^k K(u)\,\mathrm{d}u]^2}\right)^{\frac{1}{2k+1}} n^{-\frac{1}{2k+1}}.$$

This result also can be obtained by applying part (i) of the following theorem, which establishes the asymptotic MSE and MISE representations, as derived in Appendix C.

Theorem 3 *Under the assumptions given in Appendices B and C:*

(i) *We have for any fixed* $0 < t < \tau_L$*:*

$$E[\widehat{\lambda}_n(t)-\lambda(t)]^2=h_n^{2k}\left(\frac{\lambda^{(k)}(t)}{k!}\int u^k K(u)du\right)^2+\frac{\sigma^2(t)}{nh_n}+o((nh_n)^{-1})+o(h_n^{2k}).$$

(ii) *If* $\int \lambda(t)w(t)/[1-L(t)]dt < \infty$ *and* $\int l(t)w(t)/[1-L(t)]^2 dt < \infty$*, we have:*

$$E\int[\widehat{\lambda}_n(t)-\lambda(t)]^2 w(t)dt = h_n^{2k}\left(\int[\lambda^{(k)}(t)/k!]^2 w(t)dt\right)\left(\int u^k K(u)du\right)^2 +(nh_n)^{-1}\int\sigma^2(t)w(t)dt+o((nh_n)^{-1})+o(h_n^{2k}),$$

where $w(t)$ *is a weight function used to eliminate endpoint effects. A typical weight function is* $w(t) = 1$ *for* t *in some interval, say* $[\tau_L, \tau_U]$*, and* $w(t) = 0$ *otherwise.*

As a consequence of Theorem 3(ii), the asymptotically optimal bandwidth that minimizes the asymptotic mean integrated squared error is

$$h_{opt} = \left(\frac{\int\sigma^2(t)w(t)\mathrm{d}t}{\left[\int(\lambda^{(k)}(t)/k!)^2 w(t)\mathrm{d}t\right]\left[\int u^k K(u)\mathrm{d}u\right]^2}\right)^{\frac{1}{2k+1}} n^{-\frac{1}{2k+1}}. \tag{9}$$

Obviously, h_{opt} depends on the unknowns $\sigma^2(t)$ and $\lambda^{(k)}(t)$. One simple and direct way to obtain an estimator of h_{opt}, say $\widehat{h}_{\mathrm{opt}}$, is to substitute consistent estimators of $\sigma^2(t)$ and $\lambda^{(k)}(t)$, say $\widehat{\sigma}_n^2(t)$ and $\widehat{\lambda}_n^{(k)}(t)$, into (9). Clearly, $\widehat{h}_{\mathrm{opt}}$ defines a consistent estimator of h_{opt}, but we need another bandwidth selection procedure because $\widehat{h}_{\mathrm{opt}}$ depends on bandwidths through $\widehat{\sigma}_n^2(t)$ and $\widehat{\lambda}_n^{(k)}(t)$. Further research is required to assess the effect of bandwidth selection on $\widehat{h}_{\mathrm{opt}}$ and to investigate its asymptotic optimality in the context of $\widehat{h}_{\mathrm{opt}}$ asymptotically minimizing the integrated weighted squared error. Theoretical research on this problem is beyond the scope of this paper, but the next section discusses one data-driven approach to selecting bandwidths. We conclude this section with the following remark.

Remark 1 Theorem 3 shows that the choice of bandwidths b_n and γ_n does not affect the first-order term of the mean squared error, though it might affect higher order terms. Consequently, the selection of b_n and γ_n is not critical to the estimator $\widehat{\lambda}_n(t)$, a result which is also verified in our simulation study. Thus, in the next section, we consider the selection of h_n only.

4 Data-driven bandwidth selection

Cross-validation techniques based on least squares regression have been applied to density estimation for censored data by Marron and Padgett (1987). These techniques were extended to hazard estimation by Sarda and Vieu (1991) and Patil (1993a) for uncensored data, and by Patil (1993b) and González-Manteiga et al. (1996) for censored data. We further extend the least squares cross-validation approach to the case of hazard estimation for censored data with missing censoring indicators.

In our situation, the least squares cross-validated bandwidth is the minimizer of

$$CV(h_n) = \int \widehat{\lambda}_n^2(t)w(t)\mathrm{d}t - \frac{2}{n}\sum_{i=1}^{n}\frac{\widehat{\lambda}_n^{(-i)}(X_i)}{1-L_n(X_i)}w(X_i)Q_n(X_i,\delta_i,\xi_i), \quad (10)$$

where $\widehat{\lambda}_n(t)$ is one of the estimators in (5)–(7) and $\widehat{\lambda}_n^{(-i)}(t)$ is the "leave-one-out" version of that estimator. Here $Q_n(X_i,\delta_i,\xi_i)$ denotes $m_n(X_i)$, $\xi_i\delta_i + (1-\xi_i)m_n(X_i)$, or $\xi_i\delta_i/\pi_n(X_i) + [1-\xi_i/\pi_n(X_i)]m_n(X_i)$, respectively, according to whether $\widehat{\lambda}_n(t)$ represents $\widehat{\lambda}_{n,S}(t)$, $\widehat{\lambda}_{n,I}(t)$, or $\widehat{\lambda}_{n,W}(t)$. Define the cross-validated bandwidth $h_{\mathrm{opt},n}$ to be the minimizer of the score function in (10). Similar to Patil (1993a), one can establish asymptotic optimality, in the sense that

$$\frac{\mathrm{ISE}_w(h_{\mathrm{opt},n})}{\inf_{h_n\in\mathcal{H}_n}\mathrm{ISE}_w(h_n)} \xrightarrow{a.s.} 1,$$

where $\mathcal{H}_n$ is a set of bandwidths satisfying certain regularity conditions and ISE_w is the integrated weighted squared error: $\mathrm{ISE}_w(h) = \int[\widehat{\lambda}_n(t)-\lambda(t)]^2w(t)\mathrm{d}t$.

5 Simulation study

We conducted a simulation study to compare the finite sample properties of our estimators with those of the complete case estimator, say $\widehat{\lambda}_{CC}(t)$, and the standard estimator $\lambda_n(t)$ in (1). Recall that $\widehat{\lambda}_{CC}(t)$ is obtained by applying formula (1) to the subset of the data for which we observe δ. In practice we cannot compute $\lambda_n(t)$ if any of the censoring indicators are missing, but each δ is known in a simulation and thus we use $\lambda_n(t)$ as a "gold standard" for our comparisons. We now report the results of our simulation, conducted using Fortran and R programs.

The Weibull distribution is very flexible and is often used to analyze lifetime data. Thus, we generated the failure time T and censoring time C from a Weibull distribution with shape parameter τ and scale parameter η, denoted by $W(\tau,\eta)$. Given T and C, we defined $X = \min(T,C)$ and $\delta = I(T \le C)$ for each subject. We fixed $(\tau,\eta) = (3,1)$ for T, and we specified $(\tau,\eta) = (2,1.96)$ for C to obtain a 20% censoring rate, $(\tau,\eta) = (2,1.25)$ for C to obtain a 40% censoring rate, and $(\tau,\eta) = (2,0.74)$ for C to obtain a 70% censoring rate. We used the logistic model $\pi(x) = [1+\exp(-\theta_1-\theta_2x)]^{-1}$ to classify some of the censoring indicators as missing. Given $X = x$, the missingness indicator ξ was set to 1 with probability $\pi(x)$; otherwise ξ was set to 0 (and δ was treated as missing). We denoted $\pi(x)$ by $\pi_1(x)$ or $\pi_2(x)$ when the corresponding average missingness rate was approximately 20 or 40%, respectively. When the censoring rate (CR) was 20%, we set (θ_1,θ_2) to $(0.7,0.87)$ for $\pi_1(x)$ and $(0.32,0.1)$ for $\pi_2(x)$. Similarly, for CR $=$ 40%, we set (θ_1,θ_2) to $(0.7,0.98)$ for $\pi_1(x)$ and $(0.33,0.1)$ for $\pi_2(x)$; and for CR $=$ 70%, we set (θ_1,θ_2) to $(0.7,1.28)$ for $\pi_1(x)$ and $(0.33,0.13)$ for $\pi_2(x)$. We generated 1,000 samples of size

$n = 30$, 60, and 120 for each choice of CR and $\pi(x)$. We used the biweight kernel function $K(u) = \frac{15}{16}(1-u^2)^2$ if $|u| \le 1$ and $K(u) = 0$ otherwise, and the uniform kernel functions $W(u) = \Omega(u) = \frac{1}{2}$ if $|u| \le 1$ and $W(u) = \Omega(u) = 0$ otherwise. Finally, we took $w(t) = 1$ if $t \in [0, 2]$, and 0 otherwise.

First, we investigated how the bandwidth h_n obtained via least squares cross-validation varied with bandwidths b_n and γ_n. Note that $\widehat{\lambda}_{n,W}(t)$ depends on both b_n and γ_n, whereas $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,I}(t)$ depend only on b_n, and $\widehat{\lambda}_{CC}(t)$ and $\lambda_n(t)$ do not depend on either b_n or γ_n. A cross-validation bandwidth $h_{opt,n}$, the minimizer of $CV(h_n)$ in (10), was calculated for each pair (b_n, γ_n) in a 50×50 grid on the $(0, 1] \times (0, 1]$ plane. For every combination of n, CR, and $\pi(x)$, this was repeated for all estimators in each of the 1,000 samples; then the corresponding 1,000 values of $h_{\text{opt},n}$ were averaged. Figure 1a shows the results for $\widehat{\lambda}_{n,W}(t)$ when $n = 60$,

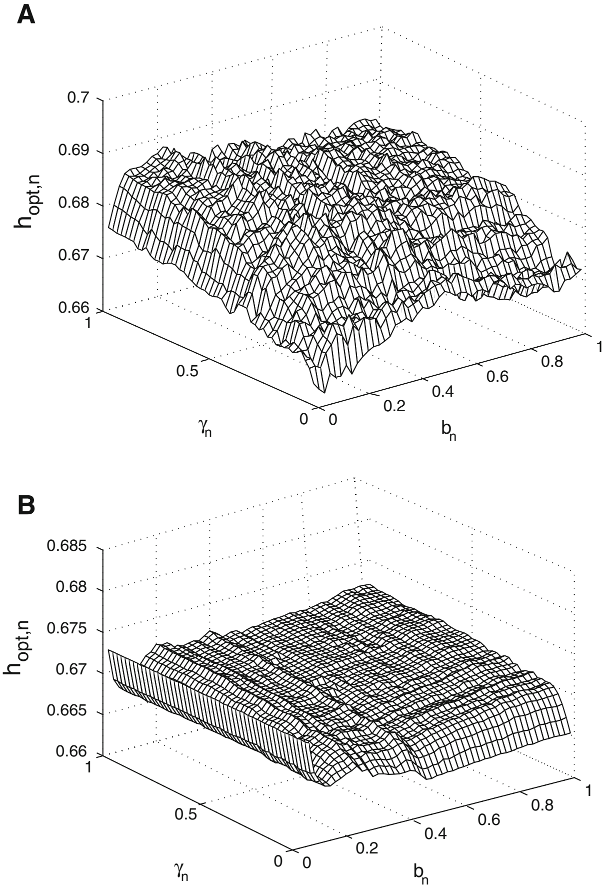

Fig. 1 **a** CV-Optimal bandwidth $h_{\text{opt},n}$ against bandwidths b_n and γ_n. The $h_{\text{opt},n}$ surface is the average $h_{\text{opt},n}$ over 1,000 replicate samples of size $n = 60$, with a censoring rate of CR $= 20\%$ and a missingness rate given by $\pi_2(x)$. These results are for $\widehat{\lambda}_{n,W}(t)$. **b** MISE-optimal bandwidth $h_{\text{opt},n}$ against bandwidths b_n and γ_n. The $h_{\text{opt},n}$ surface is the average $h_{\text{opt},n}$ over 1,000 replicate samples of size $n = 60$, with a censoring rate of CR = 20% and a missingness rate given by $\pi_2(x)$. These results are for $\widehat{\lambda}_{n,W}(t)$

CR = 20%, and $\pi(x) = \pi_2(x)$. The average values of $h_{\text{opt},n}$ ranged from 0.6616 to 0.6899, indicating that b_n and γ_n had little effect on the optimal choice of bandwidth h_n for $\widehat{\lambda}_{n,W}(t)$. Similarly, the optimal choice of h_n for $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,I}(t)$ did not vary much with b_n. These results were consistent across different values of n, CR, and $\pi(x)$.

Alternatively, an optimal bandwidth h_n can be obtained by minimizing the mean integrated squared error; refer to the left part of the equation in Theorem 3(ii) and notice that the true hazard function is known to be $\lambda(t) = 3t^2$ in our study. We also investigated how this MISE-bandwidth h_n varied with respect to (b_n, γ_n), each pair of which came from a 50×50 grid on the $(0, 1] \times (0, 1]$ plane. For every combination of n, CR, and $\pi(x)$, this was repeated for all estimators in each of the 1,000 samples; then the corresponding 1,000 values of $h_{opt,n}$ were averaged. Figure 1b shows the results for $\widehat{\lambda}_{n,W}(t)$ when $n = 60$, CR = 20%, and $\pi(x) = \pi_2(x)$. The average values of $h_{opt,n}$ ranged from 0.6642 to 0.6727, indicating that b_n and γ_n had little effect on the optimal choice of bandwidth h_n for $\widehat{\lambda}_{n,W}(t)$.

Second, we studied the effects of b_n and γ_n on MISE. We set h_n equal to the theoretically MISE-optimal bandwidth corresponding to $(b_n, \gamma_n) = (n^{-1/3}, n^{-1/3})$. For each combination of n, CR, and $\pi(x)$, we generated 1,000 samples and calculated the MISE from 1,000 simulated values of the five estimators for each pair (b_n, γ_n) from a 50×50 grid on the $(0, 1] \times (0, 1]$ plane. Figure 2 shows the results for $\widehat{\lambda}_{n,W}(t)$ when $n = 60$, CR = 40%, $\pi(x) = \pi_2(x)$, and $h_{MISE} = 0.6704$. The MISE values in Fig. 2 ranged from 1.5763 to 1.8274, which shows that the choices of b_n and γ_n were not critical for $\widehat{\lambda}_{n,W}(t)$. Similarly, the choice of b_n in our simulation had little impact for $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,I}(t)$. These results support Remark 1.

Next, Table 1 gives the MISE values for all estimators and every combination of n, CR, and $\pi(x)$. We used $(b_n, \gamma_n) = (n^{-\frac{1}{3}}, n^{-\frac{1}{3}})$, and we set h_n equal to the MISE-optimal bandwidth. In most cases, regardless of the choice of n, CR, and $\pi(x)$, the

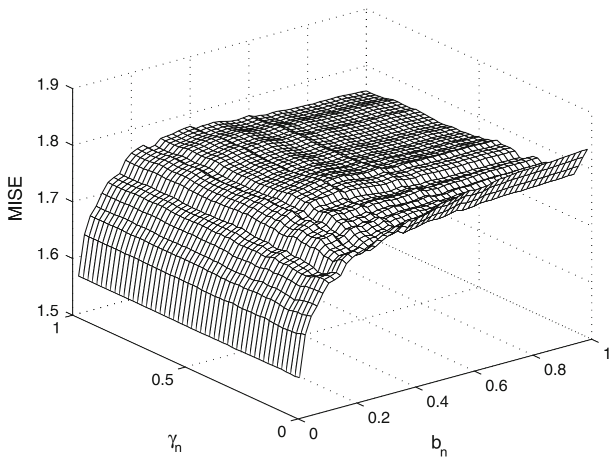

Fig. 2 Mean integrated squared error as a function of bandwidths b_n and γ_n. The MISE surface is calculated from 1,000 simulated values of $\widehat{\lambda}_{n,W}(t)$, with $n = 60$, a censoring rate of CR = 20%, a missingness rate given by $\pi_2(x)$, and a bandwidth of $h_n = 0.6693$. These results are for the inverse probability weighted estimator $\widehat{\lambda}_{n,W}(t)$

Table 1 Mean integrated squared error (MISE) by sample size (n), censoring rate (CR), and missingness rate (π) for five hazard estimators

n	Estimators	CR = 20%		CR = 40%		CR = 70%	
		$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$	$\pi_1(x)$	$\pi_2(x)$
30	λ_n	1.8190	1.9377	2.7336	2.7266	8.2347	8.1571
	$\widehat{\lambda}_{n,S}$	1.9690	2.1791	2.9608	3.0690	8.3381	10.6707
	$\widehat{\lambda}_{n,I}$	1.9705	2.1508	2.8103	2.9978	8.7351	11.3675
	$\widehat{\lambda}_{n,W}$	1.9889	2.1740	2.8054	3.0194	8.8184	11.7497
	$\widehat{\lambda}_{CC}$	2.8129	3.9094	3.6486	6.5770	9.5875	12.7147
60	λ_n	0.9218	0.9405	1.4787	1.5136	6.6776	6.6125
	$\widehat{\lambda}_{n,S}$	0.9723	1.1282	1.6166	1.6238	7.3996	10.7629
	$\widehat{\lambda}_{n,I}$	0.9694	1.1125	1.5496	1.5530	7.5676	11.1837
	$\widehat{\lambda}_{n,W}$	0.9728	1.1337	1.5480	1.5508	7.6300	11.4496
	$\widehat{\lambda}_{CC}$	1.6155	3.1192	2.2758	4.8462	8.9250	12.8148
120	λ_n	0.5241	0.5431	0.9289	0.9148	5.0455	5.0376
	$\widehat{\lambda}_{n,S}$	0.5221	0.6662	0.9866	1.0507	6.0436	10.3842
	$\widehat{\lambda}_{n,I}$	0.5424	0.6936	0.9927	1.0698	6.2026	10.6587
	$\widehat{\lambda}_{n,W}$	0.5503	0.7220	1.0051	1.1281	6.2808	10.8424
	$\widehat{\lambda}_{CC}$	1.0394	2.5448	1.5864	4.1596	8.1533	12.7515

Each entry is calculated from 1,000 simulated values of the corresponding estimator

MISE values for the three proposed estimators were fairly similar to each other and were usually close to the MISE for the gold standard, $\lambda_n(t)$. As expected, the MISE values decreased (i.e., performance improved) as the sample size increased and as the censoring percentage decreased. Except for $\lambda_n(t)$, which always used the complete censoring information, the MISE values also decreased as the missingness rate decreased (as expected). The MISE values for $\lambda_n(t)$ differed slightly from $\pi_1(x)$ to $\pi_2(x)$ simply as a result of random fluctuation in the data generated in the two sets of 1000 samples. In every situation, all three of our proposed estimators performed better than the complete case estimator, as illustrated by their smaller MISE values.

Finally, plots of the true hazard rate and average curves associated with all five estimators are presented in Fig. 3 for samples of size $n = 60$. In each of the four subplots, which correspond to the four combinations of $\pi(x)$ and the lower two censoring rates, the dotted line shows the true hazard rate $\lambda(t) = 3t^2$. The other curves are time-specific averages of the 1000 hazard estimates corresponding to $\lambda_n(t)$, $\widehat{\lambda}_{n,S}(t)$, $\widehat{\lambda}_{n,I}(t)$, $\widehat{\lambda}_{n,W}(t)$, and $\widehat{\lambda}_{CC}(t)$. Overall, the complete case estimator is the worst of the five estimators, as should be expected when a substantial amount of data is ignored. The proposed estimators lead to average curves that are nearly identical to each other and to the gold standard $\lambda_n(t)$. It is very interesting that the so-called "gold standard" is not much better than the proposed estimators with missing cause-of-death. As pointed out by one of the referees, one reason could be the improvement attributable to the "presmoothing" already observed in the literature; see, e.g., Cao and Jácome (2004) and Cao et al. (2005). As the amount of censoring increases, the proposed estimators

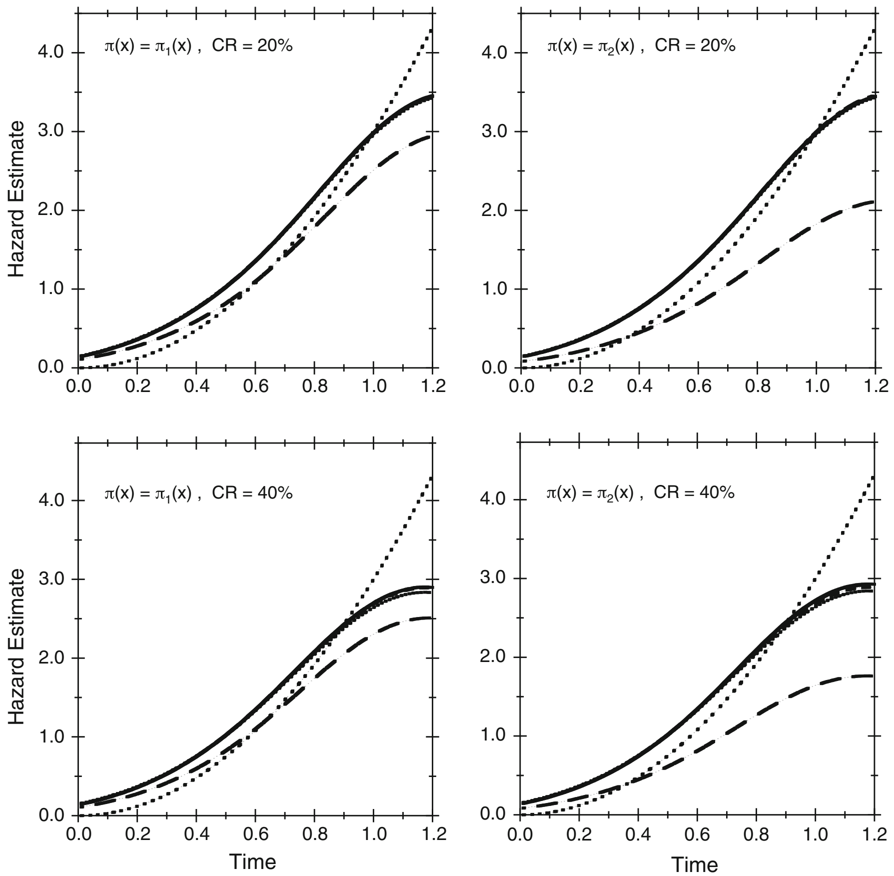

Fig. 3 True hazard rate and average curves for five hazard estimators. The true hazard function, $\lambda(t) = 3t^2$, is shown by the *wide-dot line*. The other curves are time-specific averages of estimates based on 1,000 replicate samples of size $n = 60$. We depict $\lambda_n(t)$, $\widehat{\lambda}_{n,S}(t)$, $\widehat{\lambda}_{n,I}(t)$, $\widehat{\lambda}_{n,W}(t)$, and $\widehat{\lambda}_{CC}(t)$ by *solid*, *small-dot*, *short-dash*, *dot-dash*, and *long-dash lines*, respectively. For bandwidths, we use $b_n = \gamma_n = n^{-1/3}$ and h_n is the minimizer of $CV(h_n)$. The plots on the left correspond to $\pi(x) = \pi_1(x)$ and those on the right to $\pi(x) = \pi_2(x)$. The *upper plots* correspond to 20% censoring and the *lower plots* to 40% censoring

diverge from the true curve at the largest times, but they are not nearly as biased as the complete case estimator. Also, the average curves for the proposed estimators are not affected by an increase in the amount of missingness, but the bias of the complete case estimator increases appreciably in these situations.

6 Vascular disease application

This section illustrates our methods by applying them to some data from an animal experiment. These data were previously analyzed by Dinse (1986), who reported the survival time and disease status at death for 58 female mice. At necropsy, each mouse was examined for non-renal vascular disease (NRVD). Survival was measured in days

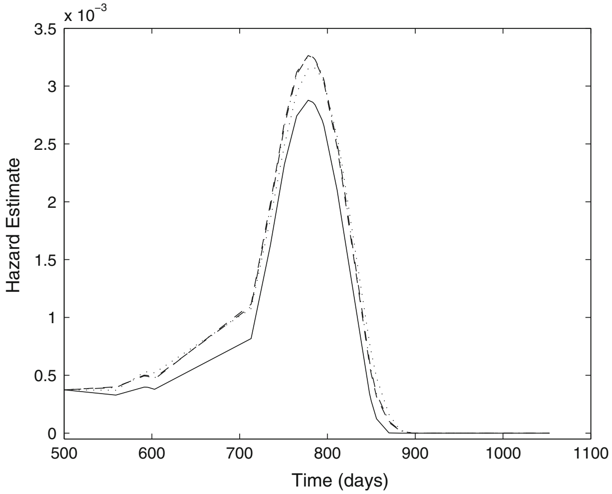

Fig. 4 Hazard estimates for death due to non-renal vascular disease (NRVD) for 33 female mice with NRVD. We depict $\widehat{\lambda}_{n,S}(t)$, $\widehat{\lambda}_{n,I}(t)$, $\widehat{\lambda}_{n,W}(t)$, and $\widehat{\lambda}_{CC}(t)$ by *dotted*, *dot-dash*, *dashed*, and *solid lines*, respectively. For bandwidths, we use $b_n = \gamma_n = 100n^{-\frac{1}{3}} \cong 31$ days and h_n is the minimizer of $CV(h_n)$

and NRVD status at death was classified as absent, incidental, unknown, or fatal. An occurrence of NRVD was considered incidental if it was present but not responsible for death and fatal if the mouse died as a direct or indirect result of its disease. In some cases, NRVD was found to be present, but its role in causing death was unknown.

We applied our methods to estimate the hazard function for death due to NRVD, say $\lambda(t)$, among the subset of mice with the disease. Time to death (X) was known for all mice. We used the same kernel functions and cross-validation bandwidth selection method as in Sect. 5. Also, we used $w(t) = I(X_{(1)} < t < X_{(n)})$, where $X_{(1)} = \min\{X_1, X_2, \ldots, X_n\}$ and $X_{(n)} = \max\{X_1, X_2, \ldots, X_n\}$. Of the 33 mice that died with NRVD present, 8 died from their disease $(\delta = 1)$, 19 died from other known causes $(\delta = 0)$, and 6 had an unknown cause of death (δ missing). Thus, 18% of the mice had a missing censoring indicator, and among mice with a known cause of death, 70% of the death times were censored with respect to the cause of interest. Figure 4 displays estimates of $\lambda(t)$ based on the proposed kernel estimators and the complete case estimator. All four curves are bell shaped and peak between 750 and 800 days. The three proposed hazard estimates are nearly identical to each other, whereas the complete case estimate is smaller, which is consistent with the simulation results plotted in Fig. 3.

7 Concluding remarks

Theoretically, all three proposed estimators have the same asymptotic representation (see Lemma 1 in Appendix B) and hence they all have the same asymptotic normal distribution (see Theorem 2). This asymptotic equivalence is similar to that obtained by Cheng (1994) for the marginal average estimator and the imputation estimator in the nonparametric regression context, and to that obtained by Wang et al. (2004) for the marginal average estimator, the regression imputation estimator, and the marginal propensity score weighted estimator in the semi-parametric regression context. Furthermore, it is shown that our estimators have the same asymptotic MSE and MISE representations (see Theorem 3). In small samples, however, simulation results show that $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,I}(t)$ have slightly smaller MISE values than $\widehat{\lambda}_{n,W}(t)$. In addition, unlike $\widehat{\lambda}_{n,W}(t)$, neither $\widehat{\lambda}_{n,S}(t)$ nor $\widehat{\lambda}_{n,I}(t)$ depends on estimating $\pi(x)$. Hence, for these reasons, one may prefer $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,I}(t)$ to $\widehat{\lambda}_{W}(t)$.

On the other hand, the inverse probability weighted approach enjoys the so-called "double robustness" property (see Scharfstein et al. 1999). That is, if $m(x)$ and $\pi(x)$ are specified by parametric models $m(x|\theta)$ and $\pi(x|\beta)$, respectively, the corresponding weighted estimator is consistent as long as one of the two models is specified correctly, where θ and β are finite dimensional parameters. This property implies that the weighted estimator is consistent if either $m(x)$ or $\pi(x)$ is estimated non-parametrically and the other is specified to be a known function, regardless of whether the specification is correct or not. However, the efficiency of the weighted estimator depends on the bias between the specified model and the true one; the larger the bias, the larger the loss of efficiency. The regression surrogate and imputation methods do not share this property.

Appendix A: Proof of Theorem 1

We begin by making the following assumptions:

- (A.$m\pi$): $m(\cdot)$ and $\pi(\cdot)$ are uniformly continuous functions.
- (A.$l\lambda$): $l(\cdot)$ and $\lambda(\cdot)$ are continuous functions.
- (A.K): $K(\cdot)$ is a probability density kernel function with bounded support and bounded variation.
- (A.$W\Omega$): $W(\cdot)$ and $\Omega(\cdot)$ are bounded kernel functions with bounded support.
- (A.h_n): $h_n \to 0$ and $h_n^{-1} n^{-\frac{1}{2}} \sqrt{\log\log n} \to 0$.
- (A.b_n): $b_n \to 0$ and $nb_n \to \infty$.
- (A.γ_n): $\gamma_n \to 0$ and $n\gamma_n \to \infty$.

Proof First we prove that Theorem 1 is true for $\widehat{\lambda}_{n,I}(t)$. Similar arguments prove that Theorem 1 is also true for $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,W}(t)$. Initially we write

$$\widehat{\lambda}_{n,I}(t) - \lambda(t) = \left[\frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \frac{\xi_i \delta_i + (1-\xi_i) m(X_i)}{n - R_i + 1} - \lambda(t) \right]$$
$$+ \left[\frac{1}{h_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \left(\frac{\xi_i \delta_i + (1-\xi_i) m_n(X_i)}{n - R_i + 1} - \frac{\xi_i \delta_i + (1-\xi_i) m(X_i)}{n - R_i + 1} \right) \right]$$
$$:= T_{n1}(t) + T_{n2}(t). \tag{11}$$

As $n[1 - L_n(X_i-)] = n - R_i + 1$, the first term in (11) can be rewritten as

$$\begin{aligned} T_{n1}(t) &= \left[\frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\delta_i}{n - R_i + 1} - \lambda(t)\right] \\ &\quad + \left[\frac{1}{nh_n}\sum_{i=1}^{n}\frac{(\xi_i - 1)[\delta_i - m(X_i)]}{1 - L_n(X_i-)}K\left(\frac{t - X_i}{h_n}\right)\right] \\ &:= T_{n11}(t) + T_{n12}(t). \end{aligned} \tag{12}$$

Based on the results of Wang (1999), by (A.$l\lambda$), (A.K) and (A.h_n) we have for the first term in (12):

$$\sup_{0 \le t \le \tau} |T_{n11}(t)| \xrightarrow{a.s.} 0. \tag{13}$$

The second term in (12) can be rewritten as

$$\begin{aligned} T_{n12}(t) &= \left[\frac{1}{nh_n}\sum_{i=1}^{n}\frac{(\xi_i - 1)[\delta_i - m(X_i)]}{1 - L(X_i)}K\left(\frac{t - X_i}{h_n}\right)\right] \\ &\quad + \left[\frac{1}{nh_n}\sum_{i=1}^{n}\frac{(\xi_i - 1)[\delta_i - m(X_i)][L_n(X_i-) - L(X_i)]}{[1 - L_n(X_i-)][1 - L(X_i)]}K\left(\frac{t - X_i}{h_n}\right)\right] \\ &:= T^{[1]}_{n12}(t) + T^{[2]}_{n12}(t). \end{aligned} \tag{14}$$

Define the following shorthand notation: $L_{n1}(t) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \le t, \delta_i = 1]$, $\widetilde{L}_{n1}(t) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \le t, \xi_i = 1]$, $L_{n11}(t) = \frac{1}{n}\sum_{i=1}^{n} I[X_i \le t, \delta_i = 1, \xi_i = 1]$, $\widetilde{L}_1(t) = P(X \le t, \xi = 1)$ and $L_{11}(t) = P(X \le t, \delta = 1, \xi = 1)$. The first term in (14) can be written

$$T^{[1]}_{n12}(t) = \frac{1}{h_n}\int K\left(\frac{t - s}{h_n}\right)\left[\frac{\mathrm{d}L_{n11}(s) - m(s)\mathrm{d}\widetilde{L}_{n1}(s) - \mathrm{d}L_{n1}(s) + m(s)dL_n(s)}{1 - L(s)}\right].$$

As $\mathrm{d}L_1(s) = m(s)\mathrm{d}L(s)$ by definition, and $\mathrm{d}L_{11}(s) = m(s)\mathrm{d}\widetilde{L}_1(s)$ under the MAR assumption, this same term also can be rewritten as

$$\begin{aligned} T^{[1]}_{n12}(t) &= \frac{1}{h_n}\int K\left(\frac{t - s}{h_n}\right)\left[\frac{\mathrm{d}[L_{n11}(s) - L_{11}(s)]}{1 - L(s)} - \frac{m(s)\mathrm{d}[\widetilde{L}_{n1}(s) - \widetilde{L}_1(s)]}{1 - L(s)}\right. \\ &\quad \left. - \frac{\mathrm{d}[L_{n1}(s) - L_1(s)]}{1 - L(s)} + \frac{m(s)\mathrm{d}[L_n(s) - L(s)]}{1 - L(s)}\right]. \end{aligned} \tag{15}$$

Denote by $L^*(\cdot)$ the distribution or sub-distribution function $L(\cdot)$, $L_1(\cdot)$, $\widetilde{L}_1(\cdot)$, or $L_{11}(\cdot)$; and denote by $L^*_n(\cdot)$ the corresponding empirical function $L_n(\cdot)$, $L_{n1}(\cdot)$, $\widetilde{L}_{n1}(\cdot)$, or $L_{n11}(\cdot)$. Using the law of the iterated logarithm for empirical distribution and sub-distribution functions: $\sup_{0\le s<\infty} |L_n^*(s) - L^*(s)| = O(n^{-\frac{1}{2}}\sqrt{\log\log n})$, $a.s.$, integration by parts, and the bounded variation of K, it follows that

$$\sup_{0\le t\le\tau} |T_{n12}^{[1]}(t)| = O(h_n^{-1} n^{-\frac{1}{2}}\sqrt{\log\log n})\ ,\ a.s. \tag{16}$$

With respect to $T_{n12}^{[2]}(t)$, similar to (16) it can be shown that $\sup_{0\le t\le\tau} |T_{n12}^{[2]}(t)| \xrightarrow{a.s.} 0$. This together with (16) proves that $\sup_{0\le t\le\tau} |T_{n12}(t)| \xrightarrow{a.s.} 0$, and thus, together with (13), we have $\sup_{0\le t\le\tau} |T_{n1}(t)| \xrightarrow{a.s.} 0$. To prove Theorem 1, it remains to prove that $\sup_{0\le t\le\tau} |T_{n2}(t)| \overset{a.s.}{\to} 0$. First, we rewrite $T_{n2}(t)$ as

$$T_{n2}(t) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{(1-\xi_i)[m_n(X_i)-m(X_i)]}{1-L_n(X_i-)}.$$

Then, by (A.$l\lambda$), (A.$m\pi$), (A.WΩ) and (A.b_n), we have

$$\sup_{0\le t\le\tau} |T_{n2}(t)| \le \sup_{0\le x\le\tau} |m_n(x)-m(x)| \sup_{0\le t\le\tau}\left|\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{1}{1-L_n(X_i-)}\right| \tag{17}$$

This proves Theorem 1 for $\widehat{\lambda}_{n,I}(t)$ and thus also for $\widehat{\lambda}_{n,S}(t)$. By similar arguments, together with the conditions listed in Appendix A and the fact that $\sup_{0\le s\le\tau} |\pi_n(s) - \pi(s)| \xrightarrow{a.s.} 0$, we can also prove that Theorem 1 is true for $\widehat{\lambda}_{n,W}(t)$. □

Appendix B: Proof of Theorem 2

We assume that the following conditions are true:

(B.$ml\pi\lambda$): $m(\cdot)$, $l(\cdot)$, $\pi(\cdot)$, and $\lambda(\cdot)$ have bounded derivatives of order $k > 1$.
(B.K): $K(\cdot)$ is a continuous kernel function of order $k > 1$ with bounded support.
(B.WΩ): $W(\cdot)$ and $\Omega(\cdot)$ are bounded kernel functions of order $k > 1$ with bounded support.
(B.h_n): $nh_n \to \infty$, $nh_n^{2k+1} = O(1)$.
(B.h_nb_n): $b_n/h_n \to 0$.
(B.$h_n\gamma_n$): $\gamma_n/h_n \to 0$.
(B.$h_nb_n\gamma_n$): $nb_n\gamma_n/h_n \to \infty$, $nh_nb_n^{2k} \to 0$, $nh_n\gamma_n^{2k} \to 0$, $h_nb_n\gamma_n^{-\frac{1}{2}} \to 0$, $h_nb_n^{-\frac{1}{2}}\gamma_n \to 0$.

Remark 2 Conditions (B.h_n), (B.h_nb_n), (B.$h_n\gamma_n$), and (B.$h_nb_n\gamma_n$) are clearly satisfied for $k = 2$ and $h_n = O(n^{-\frac{1}{5}})$, $b_n = O(n^{-\frac{1}{3}})$, and $\gamma_n = O(n^{-\frac{1}{3}})$.

Lemma 1 *If $\widehat{\lambda}_n(t)$ denotes any one of the proposed estimators $\widehat{\lambda}_{n,S}(t)$, $\widehat{\lambda}_{n,I}(t)$, or $\widehat{\lambda}_{n,W}(t)$, then under the above assumptions, we have*

$$\widehat{\lambda}_n(t) - \lambda(t) = \frac{\widetilde{f}_n(t) - E\widetilde{f}_n(t)}{1 - L(t)} + \frac{1}{nh_n}\sum_{i=1}^{n}\frac{[\xi_i - \pi(X_i)][\delta_i - m(X_i)]}{\pi(X_i)[1 - L(X_i)]}K\left(\frac{t - X_i}{h_n}\right)$$
$$+(-1)^k h_n^k \frac{\lambda^{(k)}(t)\int u^k K(u)du}{k!} + o_p\left((nh_n)^{-\frac{1}{2}}\right),$$

where $\widetilde{f}_n(t) = \frac{1}{h_n}\int K\left(\frac{t-s}{h_n}\right) dL_{n1}(s)$.

Proof (a) First we prove that Lemma 1 is true for $\widehat{\lambda}_{n,S}(t)$. We can write

$$\begin{aligned}\widehat{\lambda}_{n,S}(t) - \lambda(t) &= \left[\frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\delta_i}{n - R_i + 1} - \lambda(t)\right]\\
&\quad + \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{m(X_i) - \delta_i}{n - R_i + 1}\\
&\quad + \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{m_n(X_i) - m(X_i)}{n - R_i + 1}\\
&:= U_{n1}(t) + U_{n2}(t) + U_{n3}(t). \qquad (18)\end{aligned}$$

The first term in (18) can be rewritten as

$$\begin{aligned}U_{n1}(t) &= \left[\frac{1}{h_n}\int K\left(\frac{t - s}{h_n}\right)\frac{\mathrm{d}\widehat{F}_n(s)}{1 - \widehat{F}_n(s)} - \frac{1}{h_n}\int K\left(\frac{t - s}{h_n}\right)\frac{\mathrm{d}F(s)}{1 - F(s)}\right]\\
&\quad + \left[\frac{1}{h_n}\int K\left(\frac{t - s}{h_n}\right)\frac{\mathrm{d}F(s)}{1 - F(s)} - \lambda(t)\right]\\
&:= U_{n11}(t) + U_{n12}(t), \qquad (19)\end{aligned}$$

where $\widehat{F}_n(t)$ is the Kaplan–Meier product-limit estimator (Kaplan and Meier 1958). Based on the results of Diehl and Stute (1988), we have

$$U_{n11}(t) = \frac{\widetilde{f}_n(t) - E\widetilde{f}_n(t)}{1 - L(t)} + O_p\left((nh_n)^{-1}\right) + O_p\left(n^{-\frac{1}{2}}\right). \qquad (20)$$

Under conditions (B.$ml\pi\lambda$) and (B.K), the second term in (19) can be written

$$\begin{aligned}U_{n12}(t) &= \frac{1}{h_n}\int K\left(\frac{t - s}{h_n}\right)\lambda(s)\mathrm{d}s - \lambda(t)\\
&= (-1)^k\frac{h_n^k\lambda^{(k)}(t)}{k!}\int u^k K(u)\mathrm{d}u + o\left(h_n^k\right). \qquad (21)\end{aligned}$$

As $n[1 - L_n(X_i-)] = n - R_i + 1$ and $\sup_s |L_n(s) - L(s)| \overset{a.s.}{\to} 0$, the second term in (18) can be written as

$$U_{n2}(t) = \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \frac{m(X_i) - \delta_i}{1 - L(X_i)} + o_p\left((nh_n)^{-\frac{1}{2}}\right), \tag{22}$$

and under conditions (B.$ml\pi\lambda$) and (B.WΩ), the third term in (18) is

$$\begin{aligned} U_{n3}(t) &= \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \frac{\frac{1}{nb_n}\sum_{j=1}^{n} \xi_j[\delta_j - m(X_j)]W\left(\frac{X_i - X_j}{b_n}\right)}{[1 - L(X_i)]\pi(X_i)l(X_i)} \\ &\quad + \frac{1}{nh_n} \sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right) \frac{\frac{1}{nb_n}\sum_{j=1}^{n} \xi_j[m(X_j) - m(X_i)]W\left(\frac{X_i - X_j}{b_n}\right)}{[1 - L(X_i)]\pi(X_i)l(X_i)} \\ &\quad + o_p\left((nh_n)^{-\frac{1}{2}}\right) \\ &:= U_{n31}(t) + U_{n32}(t) + o_p\left((nh_n)^{-\frac{1}{2}}\right). \end{aligned} \tag{23}$$

Under conditions (B.K), (B.WΩ), and (B.$h_n b_n$), the first term in (23) is

$$U_{n31}(t) = \frac{1}{nh_n} \sum_{j=1}^{n} \frac{\xi_j[\delta_j - m(X_j)]}{\pi(X_j)[1 - L(X_j)]} K\left(\frac{t - X_j}{h_n}\right) + o_p\left((nh_n)^{-\frac{1}{2}}\right). \tag{24}$$

Under conditions (B.$ml\pi\lambda$), (B.K), and (B.WΩ), and following steps similar to those in the proof of equation (17) in Wang and Rao (2002), we can show that

$$U_{n32}(t) = o_p\left((nh_n)^{-\frac{1}{2}}\right). \tag{25}$$

Taken together, Eqs. (23), (24), and (25) prove

$$U_{n3}(t) = \frac{1}{nh_n} \sum_{j=1}^{n} \frac{\xi_j[\delta_j - m(X_j)]}{\pi(X_j)[1 - L(X_j)]} K\left(\frac{t - X_j}{h_n}\right) + o_p\left((nh_n)^{-\frac{1}{2}}\right), \tag{26}$$

and Eqs. (18)–(22) and (26) prove Lemma 1 for $\widehat{\lambda}_{n,S}(t)$. □

(b) Next we prove that Lemma 1 is true for $\widehat{\lambda}_{n,I}(t)$. We start by noting that the first term in (12) equals the first term in (18):

$$T_{n11}(t) = U_{n1}(t). \tag{27}$$

Based on Eq. (14), conditions (B.K) and (B.ml$\pi\lambda$), and the MAR assumption, it is straightforward to prove

$$T_{n12}(t) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{(\xi_i - 1)[\delta_i - m(X_i)]}{1 - L(X_i)} K\left(\frac{t - X_i}{h_n}\right) + o_p\left((nh_n)^{-\frac{1}{2}}\right). \quad (28)$$

Under Eqs. (12), (19), (20), (21), (27) and (28), it follows that

$$T_{n1}(t) = \frac{\widetilde{f}_n(t) - E\widetilde{f}_n(t)}{1 - L(t)} + \frac{1}{nh_n}\sum_{i=1}^{n}\frac{(\xi_i - 1)[\delta_i - m(X_i)]}{1 - L(X_i)} K\left(\frac{t - X_i}{h_n}\right)$$
$$+(-1)^k \frac{h_n^k \lambda^{(k)}(t)}{k!}\int u^k K(u)\mathrm{d}u + o_p\left((nh_n)^{-\frac{1}{2}}\right). \quad (29)$$

Similar to Eq. (26), we can show that

$$T_{n2}(t) = \frac{1}{nh_n}\sum_{i=1}^{n}\frac{\xi_i[1 - \pi(X_i)][\delta_i - m(X_i)]}{\pi(X_i)[1 - L(X_i)]} K\left(\frac{t - X_i}{h_n}\right) + o_p\left((nh_n)^{-\frac{1}{2}}\right). \quad (30)$$

Together Eqs. (11), (29), and (30) prove that Lemma 1 is true for $\widehat{\lambda}_{n,I}(t)$. $\square$

c) Finally, we prove that Lemma 1 is true for $\widehat{\lambda}_{n,W}(t)$. First note that

$$\begin{aligned}
\widehat{\lambda}_{n,W}(t) - \lambda(t) \;=\; & \left[\frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\xi_i\delta_i/\pi_n(X_i) + [1 - \xi_i/\pi_n(X_i)]m_n(X_i)}{n - R_i + 1}\right.\\
& \left. - \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\xi_i\delta_i/\pi_n(X_i) + [1 - \xi_i/\pi_n(X_i)]m(X_i)}{n - R_i + 1}\right]\\
& + \left[\frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\xi_i\delta_i/\pi_n(X_i) + [1 - \xi_i/\pi_n(X_i)]m(X_i)}{n - R_i + 1}\right.\\
& \left. - \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\xi_i\delta_i/\pi(X_i) + [1 - \xi_i/\pi(X_i)]m(X_i)}{n - R_i + 1}\right]\\
& + \left[\frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t - X_i}{h_n}\right)\frac{\xi_i\delta_i/\pi(X_i) + [1 - \xi_i/\pi(X_i)]m(X_i)}{n - R_i + 1} - \lambda(t)\right]\\
:=\; & R_{n1}(t) + R_{n2}(t) + R_{n3}(t).
\end{aligned} \quad (31)$$

Under conditions (B.K), (B.$ml\pi\lambda$), (B.WΩ), (B.h_nb_n), and (B.$h_nb_n\gamma_n$), we have

$$
\begin{aligned}
R_{n1}(t) &= \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{[1-\xi_i/\pi_n(X_i)][m_n(X_i)-m(X_i)]}{n-R_i+1} \\
&= \frac{1}{n}\sum_{j=1}^{n}\xi_j[\delta_j-m(X_j)]\frac{1}{h_nb_n}\int K\left(\frac{t-s}{h_n}\right)\frac{\frac{1-\pi(s)}{\pi(s)}W\left(\frac{s-X_j}{b_n}\right)}{[1-L(s)]\pi(s)}\mathrm{d}s \\
&\quad + o_p\left((nh_n)^{-\frac{1}{2}}\right) \\
&= O_p(n^{-\frac{1}{2}}) + o_p\left((nh_n)^{-\frac{1}{2}}\right) = o_p\left((nh_n)^{-\frac{1}{2}}\right).
\end{aligned} \tag{32}
$$

The second term in (31) can be rewritten as

$$
R_{n2}(t) = \frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{\xi_i[\delta_i-m(X_i)][\pi(X_i)-\pi_n(X_i)]}{[1-L_n(X_i-)]\pi(X_i)\widehat{\pi}_n(X_i)}.
$$

Using arguments similar to those used to derive (26), we can prove that

$$
R_{n2}(t) = o_p\left((nh_n)^{-\frac{1}{2}}\right). \tag{33}
$$

The third term in (31) can be rewritten as

$$
\begin{aligned}
R_{n3}(t) &= \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{m(X_i)}{n-R_i+1} - \lambda(t) \\
&\quad + \frac{1}{h_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{\xi_i[\delta_i-m(X_i)]}{\pi(X_i)(n-R_i+1)} \\
&= \frac{\widetilde{f}_n(t)-E\widetilde{f}_n(t)}{1-L(t)} + \frac{1}{nh_n}\sum_{i=1}^{n}\frac{[\xi_i-\pi(X_i)][\delta_i-m(X_i)]}{\pi(X_i)[1-L(X_i)]}K\left(\frac{t-X_i}{h_n}\right) \\
&\quad + h_n^k\frac{(-1)^k\lambda^{(k)}(t)\int u^kK(u)\mathrm{d}u}{k!} + o_p\left(n^{-\frac{1}{2}}\right).
\end{aligned} \tag{34}
$$

Equations (31)–(34) together prove that Lemma 1 is true for $\widehat{\lambda}_{n,W}(t)$. □

Proof of Theorem 2 By the Lyapounov central limit theorem, we have

$$
\sqrt{nh_n}\,\frac{\widetilde{f}_n(t)-E\widetilde{f}_n(t)}{1-L(t)} \xrightarrow{\mathcal{L}} N\left(0,\sigma_1^2(t)\right) \tag{35}
$$

and

$$
\frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{[\xi_i-\pi(X_i)][\delta_i-m(X_i)]}{\pi(X_i)[1-L(X_i)]}K\left(\frac{t-X_i}{h_n}\right) \xrightarrow{\mathcal{L}} N\left(0,\sigma_2^2(t)\right), \tag{36}
$$

where

$$\sigma_1^2(t) = \frac{\lambda(t)}{1 - L(t)} \int K^2(u)\mathrm{d}u$$

and

$$\sigma_2^2(t) = \frac{1 - \pi(t)}{\pi(t)} \frac{m(t)[1 - m(t)]l(t)}{[1 - L(t)]^2} \int K^2(u)\mathrm{d}u.$$

Under the MAR assumption, we can prove

$$Cov\left(\sqrt{nh_n}\frac{\widetilde{f}_n(t) - E\widetilde{f}_n(t)}{1 - L(t)}, \frac{1}{\sqrt{nh_n}}\sum_{i=1}^{n}\frac{[\xi_i - \pi(X_i)][\delta_i - m(X_i)]}{\pi(X_i)[1 - L(X_i)]}K\left(\frac{t - X_i}{h_n}\right)\right) = 0. \tag{37}$$

Equations (35)–(37), together with Lemma 1, prove that Theorem 2 is true. □

Appendix C: Proof of Theorem 3

We start by making the following assumption:
(C.π): $\inf_x \pi(x) > 0$.
First we prove that Theorem 3 is true for $\widehat{\lambda}_{n,I}(t)$. Similar proofs can be constructed for $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,W}(t)$. Recall the definitions of $T_{n11}(t)$, $T_{n12}(t)$, and $T_{n2}(t)$ in (11) and (12). Note that $T_{n11}(t) = \lambda_n(t) - \lambda(t)$, where $\lambda_n(t)$ is the standard kernel hazard function estimator given in (1). By Wang (1999), we have

$$\begin{aligned} ET_{n11}^2(t) = {} & \frac{1}{nh_n}\left(\frac{\lambda(t)}{1 - L(t)}\int K^2(u)\mathrm{d}u\right) + h_n^{2k}\left(\frac{\lambda^{(k)}(t)}{k!}\int u^k K(u)\mathrm{d}u\right)^2 \\ & + o\left(\frac{1}{nh_n}\right) + o\left(h_n^{2k}\right). \end{aligned} \tag{38}$$

Under the MAR assumption, we have

$$E\left(\frac{(\xi_i - 1)[\delta_i - m(X_i)]}{1 - L_n(X_i-)}K\left(\frac{t - X_i}{h_n}\right)\middle| X_1, X_2, \ldots, X_n\right) = 0,$$

which leads to the following result:

$$\begin{aligned} ET_{n12}^2(t) &= \frac{1}{n^2h_n^2}\sum_{i=1}^{n} E\left\{E\left(\left[\frac{(\xi_i - 1)[\delta_i - m(X_i)]}{1 - L_n(X_i-)}K\left(\frac{t - X_i}{h_n}\right)\right]^2\middle| X_1, X_2, \ldots, X_n\right)\right\} \\ &= \frac{1}{nh_n}\frac{[1 - \pi(t)]m(t)[1 - m(t)]l(t)}{[1 - L(t)]^2}\int K^2(u)\mathrm{d}u + o\left(\frac{1}{nh_n}\right). \end{aligned} \tag{39}$$

For $T_{n2}(t)$, we have

$$ET_{n2}^2(t) = E\left\{\frac{1}{nh_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{(1-\xi_i)}{1-L_n(X_i-)}\left[\frac{\sum_{j=1}^n \xi_j\delta_j W\left(\frac{X_i-X_j}{b_n}\right)}{\sum_{j=1}^n \xi_j W\left(\frac{X_i-X_j}{b_n}\right)} - m(X_i)\right]\right\}$$
$$= M_n(t) + o((M_n(t)), \tag{40}$$

where

$$M_n(t) = E\left[\frac{1}{nh_n}\sum_{j=1}^{n} \xi_j[\delta_j - m(X_j)]\frac{1}{nb_n}\sum_{i=1}^{n} K\left(\frac{t-X_i}{h_n}\right)\frac{(1-\xi_i)W\left(\frac{X_i-X_j}{b_n}\right)}{[1-L(X_i)]\pi(X_i)l(X_i)}\right]^2.$$

Under conditions (B.K), (B.WΩ) and (C.π), and the fact that $b_n/h_n \to 0$, it follows that

$$M_n(t) = E\left\{\left[\frac{1}{nh_n}\sum_{i=1}^{n} \xi_i[\delta_i - m(X_i)]K\left(\frac{t-X_i}{h_n}\right)\frac{1-\pi(X_i)}{[1-L(X_i)]\pi(X_i)}\right]^2\right\}$$
$$+ o\left(\frac{1}{nh_n}\right)$$
$$= \frac{1}{nh_n}\frac{[1-\pi(t)]^2 m(t)[1-m(t)]l(t)}{[1-L(t)]^2\pi(t)}\int K^2(u)\mathrm{d}u + o\left(\frac{1}{nh_n}\right). \tag{41}$$

Equations (11), (12), and (38)–(41) together prove Theorem 3 for $\widehat{\lambda}_{n,I}(t)$. The proofs for $\widehat{\lambda}_{n,S}(t)$ and $\widehat{\lambda}_{n,W}(t)$ follow along similar lines. □

Acknowledgments This research was supported by National Science Fund for Distinguished Young Scholars, National Natural Science Foundation of China (10671198), National Science Fund for Creative Research Groups, a grant from the Research Grants Council of the Hong Kong, China (HKU 7050/06P), a grant from Key Lab of Random Complex Structures and Data Science, CAS (Qihua Wang), the Intramural Research Program of the NIH, National Institute of Environmental Health Sciences (Z01-ES-102685) (Gregg Dinse), and the Intramural Research Traineeship Award (IRTA) program of the *Eunice Kennedy Shriver* National Institute of Child Health and Human Development (NICHD), the National Institutes of Health (NIH) (Chunling Liu). The authors thank Dr. Shyamal Peddada and Dr. Norman Kaplan for their comments and suggestions.

References

Blum, J. R., Susarla, V. (1980). Maximal deviation theory of density and failure rate function estimates based on censored data. In P. R. Krishniah (Ed.), *Multivariate analysis* (pp. 213–222). New York: North-Holland.

Cao, R., Jácome, M. A. (2004). Presmoothed kernel density estimator for censored data. *Journal of Nonparametric Statistics, 16*, 289–309.

Cao, R., López-de-Ullibarri, I., Janssen, P., Veraverbeke, N. (2005). Presmoothed Kaplan–Meier and Nelson–Aalen estimators. *Journal of Nonparametric Statistics, 17*, 31–56.

Cheng, P. E. (1994). Nonparametric estimation of mean functionals with data missing at random. *Journal of the American Statistical Association, 89*, 81–87.

Dewanji, A. (1992). A note on a test for competing risks with missing failure type. *Biometrika, 79*, 855–857.

Diehl, S., Stute, W. (1988). Kernel density and hazard function estimation in the presence of censoring. *Journal of Multivariate Analysis, 25*, 299–310.
Dikta, G. (1998). On semiparametric random censorship models. *Journal of Statistical Planning and Inference, 66*, 253–279.
Dinse, G. E. (1982). Nonparametric estimation for partially-complete time and type of failure data. *Biometrics, 38*, 417–431.
Dinse, G. E. (1986). Nonparametric prevalence and mortality estimators for animal experiments with incomplete cause-of-death data. *Journal of the American Statistical Association, 81*, 328–336.
Gao, G. Z., Tsaitis, A. A. (2005). Semiparametric estimators for the regression coefficients in the linear transformation competing risks model with missing cause of failing. *Biometrika, 92*, 875–891.
Goetghebeur, E. J., Ryan, L. M. (1990). A modified log rank test for competing risks with missing failure type. *Biometrika, 77*, 207–211.
Goetghebeur, E. J., Ryan, L. M. (1995). Analysis of competing risks survival data when some failure types are missing. *Biometrika, 82*, 821–833.
González-Manteiga, W., Cao, R., Marron, J. S. (1996). Bootstrap selection of the smoothing parameter in nonparametric hazard rate estimation. *Journal of the American Statistical Association, 91*, 1130–1140.
Jacome, M. A., Gijbels, I., Cao, R. (2008). Comparison of presmoothing methods in kernel density estimation under censoring. *Computational Statistics, 23*, 381–406.
Kalbfleisch, J. D., Prentice, R. L. (1980). *The statistical analysis of failure time data*. New York: Wiley.
Kaplan, E. L., Meier, P. (1958). Nonparametric estimation from incomplete observations. *Journal of the American Statistical Association, 53*, 457–481.
Klein, J. P., Moeschberger, M. L. (2003). *Survival analysis*. New York: Springer.
Lipsitz, S. R., Zhao, L. P., Molenberghs, G. (1998). A semiparametric method of multiple imputation. *Journal of the Royal Statistical Society, Series B, 60*, 127–144.
Little, R. J. A., Rubin, D. B. (1987). *Statistical analysis with missing data*. New York: Wiley.
Lo, S.-H. (1991). Estimating a survival function with incomplete cause-of-death data. *Journal of Multivariate Analysis, 39*, 217–235.
Marron, J. S., Padgett, J. W. (1987). Asymptotically optimal bandwidth selection for kernel density estimators from randomly right-censored samples. *The Annals of Statistics, 15*, 1520–1535.
McKeague, I. W., Subramanian, S. (1998). Product-limit estimators and Cox regression with missing censoring information. *Scandinavian Journal of Statistics, 25*, 589–601.
Nelson, W. (1972). Theory and applications of hazard plotting for censored failure data. *Technometrics, 14*, 945–966.
Patil, P. N. (1993a). Bandwidth choice for nonparametric hazard rate estimation. *Journal of Statistical Planning and Inference, 35*, 15–30.
Patil, P. N. (1993b). On the least squares cross-validation bandwidth in hazard rate estimation. *The Annals of Statistics, 21*, 1792–1810.
Ramlau-Hansen, H. (1983). Smoothing counting process intensities by means of kernel functions. *The Annals of Statistics, 11*, 453–466.
Regina, Y. C., John, V. R. (1985). A histogram estimator of the hazard rate with censored data. *The Annals of Statistics, 13*, 592–605.
Robins, J. M., Rotnitzky, A. (1992). Recovery of information and adjustment for dependent censoring using surrogate markers. In N. Jewell, K. Dietz, V. Farewell (Eds.), *AIDS epidemiology—methodological issues* (pp. 297–331). Boston: Birkhäuser.
Robins, J. M., Rotnitzky, A., Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association, 89*, 846–866.
Robins, J. M., Wang, N. (2000). Inference for imputation estimators. *Biometrika, 87*, 113–124.
Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York: Wiley.
Sarda, P., Vieu, P. (1991). Smoothing parameter selection in hazard estimation. *Statistics & Probability Letters, 11*, 429–434.
Scharfstein, D. O., Rotnitzky, A., Robins, J. (1999). Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). *Journal of the American Statistical Association, 94*, 1096–1146.
Subramanian, S. (2004). Asymptotically efficient estimation of a survival function in the missing censoring indicator model. *Journal of Nonparametric Statistics, 16*, 797–817.
Subramanian, S. (2006). Survival analysis for the missing censoring indicator model using kernel density estimation techniques. *Statistical Methodology, 3*, 125–136.

Tanner, M. A. (1983). A note on the variable kernel estimator of the hazard function from randomly censored data. *The Annals of Statistics, 11*, 994–998.
Tanner, M. A., Wong, W. H. (1983). The estimation of the hazard function from randomly censored data by the kernel method. *The Annals of Statistics, 11*, 989–993.
Tsiatis, A. A., Davidian, M., McNeney, B. (2002). Multiple imputation methods for testing treatment differences in survival distributions with missing cause of failure. *Biometrika, 89*, 238–244.
van der Laan, M. J., McKeague, I. W. (1998). Efficient estimation from right-censored data when failure indicators are missing at random. *The Annals of Statistics, 26*, 164–182.
Wang, Q. H. (1999). Some bounds for the error of an estimator of the hazard function with censored data. *Statistics & Probability Letters, 44*, 319–326.
Wang, Q. H., Linton, O., Härdle, W. (2004). Semiparametric regression analysis with missing response at random. *Journal of the American Statistical Association, 99*, 334–345.
Wang, Q. H., Rao, J. N. K. (2002). Empirical likelihood-based inference under imputation for missing response data. *The Annals of Statistics, 30*, 896–924.
Zhao, L. P., Lipsitz, S. R., Lew, D. (1996). Regression analysis with missing covariate data using estimating equations. *Biometrics, 52*, 1165–1182.