Published in final edited form as:

J Am Stat Assoc. 2014 January 1; 109(505): 371–383. doi:10.1080/01621459.2013.842172.

Efficient Estimation of Semiparametric Transformation Models for Two-Phase Cohort Studies

DONGLIN ZENG [Professor] and D. Y. LIN [Dennis Gillings Distinguished Professor]

Department of Biostatistics, CB#7420, University of North Carolina, Chapel Hill, NC 27599-7420

DONGLIN ZENG: dzeng@bios.unc.edu; D. Y. LIN: lin@bios.unc.edu

Abstract

Under two-phase cohort designs, such as case-cohort and nested case-control sampling, information on observed event times, event indicators, and inexpensive covariates is collected in the first phase, and the first-phase information is used to select subjects for measurements of expensive covariates in the second phase; inexpensive covariates are also used in the data analysis to control for confounding and to evaluate interactions. This paper provides efficient estimation of semiparametric transformation models for such designs, accommodating both discrete and continuous covariates and allowing inexpensive and expensive covariates to be correlated. The estimation is based on the maximization of a modified nonparametric likelihood function through a generalization of the expectation-maximization algorithm. The resulting estimators are shown to be consistent, asymptotically normal and asymptotically efficient with easily estimated variances. Simulation studies demonstrate that the asymptotic approximations are accurate in practical situations. Empirical data from Wilms' tumor studies and the Atherosclerosis Risk in Communities (ARIC) study are presented.

Keywords

Case-cohort design; EM algorithm; Kernel estimation; Nested case-control sampling; Nonparametric likelihood; Semiparametric efficiency

1. INTRODUCTION

In epidemiological cohort studies, the outcomes of primary interest, such as cancer, heart diseases and death, are relatively rare, even after many years of follow-up; therefore, the cohorts must be very large so as to yield reliable information about the effects of exposures or other covariates on the event times. The covariates of interest often involve biochemical or genetic analysis of blood specimens or extraction of detailed exposure histories and thus are prohibitively expensive to measure on large cohorts. A cost-effective strategy is to measure the covariates on all cases, i.e., the subjects who have developed the outcome of interest during the follow-up, and a subset of controls, i.e., those who have not developed the outcome of interest by the end of the study. Two sampling schemes have been commonly used to select the controls: case-cohort sampling (Prentice, 1986) selects a random subcohort of the original cohort; nested case-control sampling (Thomas, 1977) selects a small group of controls, typically between 1 and 5, for each observed event time.

In addition to the covariates that are difficult to measure, there are covariates, such as demographical and geographical factors and basic clinical variables, that are easily measured on all cohort members. Such covariates can be used as stratification variables for case-cohort or nested case-control sampling. Thus, case-cohort and nested case-control studies are regarded as two-phase studies: in the first phase, information on the event time, case-control status, and inexpensive covariates is gathered for each cohort member; in the

second phase, the information collected in the first phase is used to determine which cohort members will be measured on expensive covariates. Inexpensive covariates can also be used in regression models to adjust for confounding and to assess interactions among covariates.

The statistical literature on case-cohort studies is rather extensive. Most of the existing work is focused on the original design of Prentice (1986), under which the subcohort is selected by simple random sampling and the first-phase covariate information is not used in the sampling or the analysis. Specifically, Prentice (1986) and Self and Prentice (1988) proposed pseudo-partial likelihood functions for the proportional hazards model (Cox, 1972) by replacing the risk sets of the whole cohort in the partial likelihood function (Cox, 1975) with their subcohort counterparts. Kalbfleisch and Lawless (1988) suggested a modification that weights individuals' contributions to the partial likelihood function by their inverse probabilities of selection; Barlow (1994), Chen and Lo (1999), Borgan et al. (2000) and Chen (2001) studied related estimators. Weighted estimators were also proposed by Kong et al. (2004) and Lu and Tsiatis (2006) for the class of linear transformation models, which includes the proportional hazards model as a special case. Similar estimators were proposed by Kulich and Lin (2000), Nan et al. (2006) and Kong and Cai (2009) for the additive risk and accelerated failure time models. Weighted estimators are not asymptotically efficient. To improve efficiency, Kulich and Lin (2004), Qi et al. (2005), Breslow et al. (2009), and Luo et al. (2009) used auxiliary variables to adjust the sampling weights along the lines of Robins et al. (1994). Efficient estimation of the proportional hazards model was studied by Chen and Little (1999) and Martinussen (1999) for parametric covariate distributions, by Nan (2004) for discrete covariates and by Scheike and Martinussen (2004) for the original case-cohort design.

There are a few statistical papers on nested case-control sampling. Goldstein and Langholz (1992) established the asymptotic theory of the maximum partial likelihood estimator for the proportional hazards model. Langholz and Borgan (1995) explored a stratified nested case-control sampling method via counter-matching to improve efficiency. Scheike and Juul (2004) proposed nonparametric maximum likelihood estimation for a proportional hazards model stratified by the first-phase covariates. Zeng et al. (2006) explored applications to genetic association studies under the assumption of gene-environment independence for both nested case-control and case-cohort sampling. Cai and Zheng (2011; 2012) proposed weighted estimators for evaluating biomarker accuracy in nested case-control studies.

In this paper, we study efficient estimation of semiparametric transformation models under general two-phase designs, accommodating both discrete and continuous covariates and allowing inexpensive and expensive covariates to be correlated. Efficient estimation under such general designs has not been pursued even for the special case of the proportional hazards model. We stress the importance of using the first-phase covariates, which are available in virtually all epidemiological cohort studies, to improve the efficiency of the second-phase sampling and to control for confounding and evaluate interactions. The existing literature on efficient estimation either ignores the first-phase covariates or requires them to be discrete or independent of expensive covariates. We allow the first-phase covariates to be continuous and correlated with expensive covariates and do not parametrize the distribution of covariates. It is very difficult to deal with this general situation because the likelihood function involves the conditional density functions given continuous variables. We overcome this hurdle by incorporating kernel functions for conditional density functions into the nonparametric likelihood function and devising a semiparametric EM algorithm to maximize the modified likelihood function. We establish the consistency, weak convergence and asymptotic efficiency of the resulting estimators through novel combinations of modern empirical process theory and kernel estimation theory. The

proposed methods can drastically reduce the cost of epidemiological cohort studies while incurring little loss of statistical efficiency relative to full-cohort sampling.

2. THEORY AND METHODS

2.1. Maximum Likelihood Estimation

Let T denote the event time, $\mathbf{X}$ denote the set of expensive covariates that is measured on a subset of cohort members, $\mathbf{Z}$ denote the set of completely measured covariates that is potentially correlated with $\mathbf{X}$, and $\mathbf{W}$ denote the set of completely measured covariates that is known to be independent of $\mathbf{X}$. We specify that the cumulative hazard function of T conditional on $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$ satisfies the proportional hazards model (Cox, 1972) or the more general class of semiparametric transformation models:

$$\Lambda(t|\mathbf{X}, \mathbf{Z}, \mathbf{W}) = G \left\{ \Lambda(t) e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \right\},$$

where G is a known increasing function, $\Lambda(\cdot)$ is an unspecified positive increasing function, and β , γ and η are unknown regression parameters (Zeng and Lin, 2007). The linear predictor can be modified to include interactions between $\mathbf{X}$, $\mathbf{Z}$ and $\mathbf{W}$. We consider the class of Box-Cox transformations $G(x) = \{(1+x)^\rho - 1\}/\rho$ ($\rho \neq 0$) and the class of logarithmic transformations $G(x) = \log(1+rx)/r$ ($r \neq 0$) (Chen et al., 2002). The choices of $\rho = 1$ or $r = 0$ and $\rho = 0$ or $r = 1$ yield the proportional hazards and proportional odds models, respectively. The transformation may be determined by the AIC criterion.

In the presence of right censoring, we observe Y and Δ instead of T , where $Y = \min(T, C)$, $\Delta = I(T \leq C)$, C is the censoring time, and $I(\cdot)$ is the indicator function. For a cohort of size n , the first-phase data consist of $\mathbf{O} \equiv (Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)$ ($i = 1, \dots, n$). Let R_i indicate, by the values 1 versus 0, whether the i th subject is selected for the measurement of $\mathbf{X}$ in the second phase. Denote $\mathcal{C} = \{i : R_i = 1\}$ and $\bar{\mathcal{C}} = \{i : R_i = 0\}$. We make two basic assumptions:

- (A.1) The censoring time C is independent of T conditional on $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ for $R = 1$ and independent of T and $\mathbf{X}$ conditional on $(\mathbf{Z}, \mathbf{W})$ for $R = 0$;
- (A.2) The sampling vector $(R_1, \dots, R_n)$ is independent of $(\mathbf{X}_1, \dots, \mathbf{X}_n)$ conditional on $\mathbf{O}$.

Remark 1—(A.1) is the standard coarsening-at-random assumption. Under (A.22), the selection of a subject for the measurement of $\mathbf{X}$ may depend on any first-phase data. This assumption is clearly satisfied by the original case-cohort design, where R_i depends only on Δ_i , and by the original nested case-control design, where R_i depends on Δ_i and the risk sets.

Write $\theta = (\beta^T, \gamma^T, \eta^T)^T$, and let $P(\cdot|\cdot)$ denote a conditional density function. Because of Assumption (A.2), the sampling probabilities can be omitted from the likelihood function in estimating θ and Λ . For a subject in $\mathcal{C}$, the likelihood contribution is the density of $(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$; for a subject in $\bar{\mathcal{C}}$, the likelihood contribution is the density of $(Y, \Delta, \mathbf{Z}, \mathbf{W})$. Thus, the log-likelihood function for θ , Λ and $P(\mathbf{x}|\mathbf{z})$ takes the form

$$\sum_{i \in \mathcal{C}} \{\log P(Y_i, \Delta_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + \log P(\mathbf{X}_i | \mathbf{Z}_i)\} + \sum_{i \in \bar{\mathcal{C}}} \log \int P(Y_i, \Delta_i | \mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i) P(\mathbf{x} | \mathbf{Z}_i) d\mathbf{x}. \quad (1)$$

Under Assumption (A.1), $P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i)$ (for $i \in \bar{\mathcal{C}}$) is the product of

$$\left[\Lambda'(Y_i) e^{\beta^T \mathbf{x} + \gamma^T \mathbf{Z}_i + \eta^T \mathbf{W}_i} G' \left\{ \Lambda(Y_i) e^{\beta^T \mathbf{x} + \gamma^T \mathbf{Z}_i + \eta^T \mathbf{W}_i} \right\} \right]^{\Delta_i} \exp \left[-G \left\{ \Lambda(Y_i) e^{\beta^T \mathbf{x} + \gamma^T \mathbf{Z}_i + \eta^T \mathbf{W}_i} \right\} \right] \quad (2)$$

and a function that does not involve $\mathbf{x}$ or (θ, Λ) and thus can be factored out of the integral in (1). Here and in the sequel, $f'(x) = df(x)/dx$, and $f''(x) = d^2f(x)/dx^2$.

To maximize (1) using the nonparametric maximum likelihood estimation (Johansen, 1983), we let the estimator for Λ be a step function with jumps only at the observed Y_i with $\Delta_i = 1$ and replace (2) by

$$\left[\Lambda\{Y_i\} e^{\beta^T \mathbf{x} + \gamma^T \mathbf{Z}_i + \eta^T \mathbf{W}_i} G' \left\{ \Lambda(Y_i) e^{\beta^T \mathbf{x} + \gamma^T \mathbf{Z}_i + \eta^T \mathbf{W}_i} \right\} \right]^{\Delta_i} \exp \left[-G \left\{ \Lambda(Y_i) e^{\beta^T \mathbf{x} + \gamma^T \mathbf{Z}_i + \eta^T \mathbf{W}_i} \right\} \right],$$

where $\Lambda\{Y_i\}$ denotes the jump size of $\Lambda(\cdot)$ at Y_i . In addition, we let the estimator for $P(\mathbf{X}|\mathbf{Z})$ be a discrete probability function on the distinct observed values of $\mathbf{X}$, denoted by $\mathbf{x}_1, \dots, \mathbf{x}_m$.

Even with the above discretization, maximization of (1) is not feasible if $\mathbf{Z}$ is continuous because then only a small number of observations on $\mathbf{X}$ are associated with each observed $\mathbf{Z}$. To tackle this challenge, we estimate the conditional distribution of $\mathbf{X}$ given $\mathbf{Z}$ by maximizing a local likelihood function. That is, we approximate $\log P(\mathbf{X}_i|\mathbf{Z}_i)$ and $P(\mathbf{x}|\mathbf{Z}_i)$ in (1) by $\sum_{j=1}^n w_{ji} \log P(\mathbf{X}_i|\mathbf{Z}=\mathbf{Z}_j)$ and $\sum_{j=1}^n w_{ji} P(\mathbf{x}|\mathbf{Z}=\mathbf{Z}_j)$, respectively, where

$$w_{ji} = \frac{K\{(\mathbf{Z}_j - \mathbf{Z}_i)/a_n\}}{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z}_i)/a_n\}},$$

$K(\cdot)$ is a symmetric kernel function, and a_n is a constant. Consequently, (1) becomes

$$\sum_{i \in \mathcal{C}} \left\{ \log P(Y_i, \Delta_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + \sum_{j=1}^n w_{ji} \log P(\mathbf{X}_i | \mathbf{Z}_j) \right\} + \sum_{i \in \overline{\mathcal{C}}} \log \left\{ \int P(Y_i, \Delta_i | \mathbf{x}, \mathbf{Z}_i, \mathbf{W}_i) \sum_{j=1}^n w_{ji} P(\mathbf{x} | \mathbf{Z}_j) d\mathbf{x} \right\}.$$

Denoting the point mass of $P(\mathbf{X}|\mathbf{Z}_j)$ at $\mathbf{x}_s$ by p_{sj} , we aim to maximize the following function

$$\begin{aligned} l_n(\theta, \Lambda, p_{sj}) &= \sum_{i \in \mathcal{C}} \left\{ \log P(Y_i, \Delta_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + \sum_{s=1}^m I(\mathbf{X}_i = \mathbf{x}_s) \sum_{j=1}^n w_{ji} \log p_{sj} \right\} \\ &+ \sum_{i \in \overline{\mathcal{C}}} \log \left\{ \sum_{s=1}^m P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i) \sum_{j=1}^n w_{ji} p_{sj} \right\} \end{aligned} \quad (3)$$

under the constraints of $\sum_{s=1}^m p_{sj} = 1$ and $p_{sj} \geq 0$ ($j = 1, \dots, n$). When $\mathbf{Z}$ is discrete, we can choose a_n small enough such that $w_{ji} = I(j = i)$ and (3) reduces to the original log-likelihood function with $P(\mathbf{X} = \mathbf{x}_s | \mathbf{Z} = \mathbf{Z}_i) = p_{si}$. The above estimation procedure is feasible even if $\mathbf{X}$ is multi-dimensional because the empirical distribution function of $\mathbf{X}$ given $\mathbf{Z}$ is used.

2.2. EM Algorithm

Direct maximization of (3) is difficult and potentially unstable due to the intractable form of the second term. To make the problem more tractable, we artificially create a latent variable $\tilde{\mathbf{Z}}$ for subjects in $\bar{\mathcal{C}}$ which takes values on the observed $\mathbf{Z}_1, \dots, \mathbf{Z}_n$ and satisfies the equations $P(\tilde{\mathbf{Z}} = \mathbf{Z}_j | \mathbf{Z} = \mathbf{Z}_i, \mathbf{W}) = w_{ji}$, $P(\mathbf{X} = \mathbf{x}_s | \mathbf{Z} = \mathbf{Z}_i, \tilde{\mathbf{Z}} = \mathbf{Z}_j, \mathbf{W}) = P(\mathbf{X} = \mathbf{x}_s | \mathbf{Z} = \mathbf{Z}_j) = p_{sj}$, and $P(Y, \Delta | \mathbf{X}, \mathbf{Z}, \mathbf{W}, \tilde{\mathbf{Z}}) = P(Y, \Delta | \mathbf{X}, \mathbf{Z}, \mathbf{W})$. Then $P(\mathbf{X} = \mathbf{x}_s | \mathbf{Z} = \mathbf{Z}_i) = \sum_{j=1}^n w_{ji} p_{sj}$ for subjects in $\bar{\mathcal{C}}$. Thus, the second term in (3) is equivalent to the log-likelihood of $(Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)$ ($i \in \bar{\mathcal{C}}$) assuming that the complete data consist of $(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i, \tilde{\mathbf{Z}}_i)$ ($i \in \bar{\mathcal{C}}$) but both $\mathbf{X}_i$ and $\tilde{\mathbf{Z}}_i$ are missing.

We now devise an EM-type algorithm to maximize (3) by treating $(\mathbf{X}_i, \tilde{\mathbf{Z}}_i)$ ($i \in \bar{\mathcal{C}}$) as missing. The complete-data log-likelihood for subjects in $\bar{\mathcal{C}}$ can be written as

$$\begin{aligned} & \sum_{i \in \bar{\mathcal{C}}} \left\{ \log P(Y_i, \Delta_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + \log P(\mathbf{X}_i | \tilde{\mathbf{Z}}_i) + \log P(\tilde{\mathbf{Z}}_i | \mathbf{Z}_i) + \log P(\mathbf{Z}_i, \mathbf{W}_i) \right\} \\ &= \sum_{i \in \bar{\mathcal{C}}} \left\{ \sum_{s=1}^m I(\mathbf{X}_i = \mathbf{x}_s) \log P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i) + \sum_{j=1}^n \sum_{s=1}^m I(\mathbf{X}_i = \mathbf{x}_s, \tilde{\mathbf{Z}}_i = \mathbf{Z}_j) \log p_{sj} \right. \\ & \quad \left. + \sum_{j=1}^n I(\tilde{\mathbf{Z}}_i = \mathbf{Z}_j) \log w_{ji} + \log P(\mathbf{Z}_i, \mathbf{W}_i) \right\}. \end{aligned}$$

In the E-step, we calculate the conditional expectations of $I(\mathbf{X}_i = \mathbf{x}_s)$ and $I(\mathbf{X}_i = \mathbf{x}_s, \tilde{\mathbf{Z}}_i = \mathbf{Z}_j)$ given the observed data for $i \in \bar{\mathcal{C}}$ and $j = 1, \dots, n$ as

$$\hat{q}_{is} = \frac{P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i) \sum_{j=1}^n w_{ji} p_{sj}}{\sum_{s=1}^m P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i) \sum_{j=1}^n w_{ji} p_{sj}},$$

and

$$\hat{\psi}_{sji} = \frac{w_{ji} p_{sj}}{\sum_{j=1}^n w_{ji} p_{sj}} \hat{q}_{is},$$

respectively. In the M-step, we update θ and Λ by maximizing

$$\sum_{i \in \bar{\mathcal{C}}} \log P(Y_i, \Delta_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) + \sum_{i \in \bar{\mathcal{C}}} \sum_{s=1}^m \hat{q}_{is} \log P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i). \quad (4)$$

Because (4) is a weighted sum of the log-likelihood functions under the semiparametric transformation model, the algorithms of Zeng and Lin (2007) can be modified for this optimization. We update p_{sj} ($j = 1, \dots, n; s = 1, \dots, m$) by maximizing

$$\sum_{i \in \bar{\mathcal{C}}} \sum_{s=1}^m I(\mathbf{X}_i = \mathbf{x}_s) w_{ji} \log p_{sj} + \sum_{i \in \bar{\mathcal{C}}} \sum_{s=1}^m \hat{\psi}_{sji} \log p_{sj}, \text{ so that}$$

$$p_{sj} = \frac{\sum_{i \in \mathcal{C}} I(\mathbf{X}_i = \mathbf{x}_s) w_{ji} + \sum_{i \in \mathcal{C}} \hat{\psi}_{sji}}{\sum_{s=1}^m \left\{ \sum_{i \in \mathcal{C}} I(\mathbf{X}_i = \mathbf{x}_s) w_{ij} + \sum_{i \in \mathcal{C}} \hat{\psi}_{sji} \right\}}.$$

We set the initial values of θ , $\Lambda\{\cdot\}$ and p_{sj} to $\mathbf{0}$, the inverse of the total number of cases, and m^{-1} , respectively, and iterate between the E-step and M-step until convergence. Denote the resulting estimator of θ , Λ and p_{sj} by $\hat{\theta}$, $\hat{\Lambda}$ and $\hat{p}_{sj}$. Since the maximum likelihood estimator for the distribution function of $\mathbf{Z}$ is the empirical distribution function based on $(\mathbf{Z}_1, \dots, \mathbf{Z}_n)$, the joint distribution function of $(\mathbf{X}, \mathbf{Z})$, denoted by $F(\cdot, \cdot)$, can be estimated by the distribution function $F(\cdot, \cdot)$ with point mass $\hat{p}_{sj}/n$ at $(\mathbf{x}_s, \mathbf{z}_j)$ ($s = 1, \dots, m; j = 1, \dots, n$). The proposed algorithm is not exactly an EM algorithm since the objective function given in (3) is only an approximation to a log-likelihood function. However, we show in the Supplemental Materials that the objective function increases at each iteration. The proposed algorithm may converge to a local maximum. To obtain the global maximum, we suggest to use different initial values or adopt a quasi-Newton algorithm to maximize (3).

2.3. Asymptotic Properties

Let θ_0 , Λ_0 and F_0 denote the true values of θ , Λ and F , and let τ denote the endpoint of the study. We impose the following regularity conditions:

- (C.1) The set of covariates $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ has bounded support and the joint density of $(\mathbf{X}, \mathbf{Z})$ with respect to some dominating measure is q -times continuously differentiable in the support of $(\mathbf{X}, \mathbf{Z})$, where $q > (d_{\mathbf{z}} + 3)/2$, and $d_{\mathbf{z}}$ is the dimension of $\mathbf{Z}$. If $[1, \mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T]\mathbf{v} = 0$ for some constant vector $\mathbf{v}$ with probability one, then $\mathbf{v} = \mathbf{0}$.
- (C.2) The function Λ_0 is twice continuously differentiable in $[0, \tau]$ and $\Lambda_0'(t) > 0$ for $t \in [0, \tau]$.
- (C.3) The transformation function G is three-times continuously differentiable with $G(0) = 0$, $G'(0) > 0$ and $\{1 + G'(x)\}e^{-G(x)} \sim c_1(1+x)^{-\nu_0}$ for some constants $\nu_0 > 0$ and $c_1 > 0$.
- (C.4) The kernel function $K(\cdot)$ is spherically symmetric, i.e., $K(\mathbf{z}) = \tilde{K}(\|\mathbf{z}\|)$ for some univariate and symmetric kernel function $\tilde{K}(\cdot)$. In addition, $\tilde{K}(\cdot)$ has a bounded support and satisfies the moment conditions $\int \tilde{K}(x)dx = 1$ and $\int \tilde{K}(x)x^j dx = 0$ ($j = 1, \dots, q-1$).
- (C.5) The bandwidth a_n satisfies that $a_n \rightarrow 0$, $na_n^{d_{\mathbf{z}}} \rightarrow \infty$ and $na_n^{2q} \rightarrow 0$.
- (C.6) There exists a function $p(r; y, \delta, \mathbf{z}, \mathbf{w})$ ($r = 0, 1$) such that $\sum_{r=0}^1 p(r; y, \delta, \mathbf{z}, \mathbf{w}) = 1$, $p(1; Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i) > q_0$ almost surely for some positive constant q_0 , and under the probability measure associated with $\prod_{i=1}^n p(R_i; Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)$ and conditional on $\mathbf{O}$,

$$\left\{ \log \frac{\Pr(R_1=r_1, \dots, R_n=r_n | \mathbf{O})}{\prod_{i=1}^n p(r_i; Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)} + f(\mathbf{O})/2 \right\} / f(\mathbf{O})^{1/2}$$

is asymptotically standard normal, where f is a positive and measurable function.

(C.7) With probability one, $\Pr(C \neq \emptyset | \mathbf{Z}, \mathbf{W})$ is bounded away from zero. In addition, the conditional density of C given $(\mathbf{Z}, \mathbf{W})$ is q -times continuously differentiable.

Remark 2—(C.1) is a standard condition in regression analysis, implying that $(1, \mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)$ is linearly independent with positive probability. (C.3) is satisfied by all commonly used transformation functions, including the classes of Box-Cox transformations and logarithmic transformations. The choice of the bandwidth in (C.5) is not optimal. Indeed, it is desirable to undersmooth the estimator of the conditional density $P(\mathbf{x}|\mathbf{z})$ so as to reduce the bias in the estimation of θ . We choose $a_n = n^{-\nu_2}$ with $\nu_2 \in (1/2q, 1/d_{\mathbf{z}})$. Specifically, we set $a_n = Bn^{-1/(3+d_{\mathbf{z}})}$, where B is a constant in the same magnitude as the standard deviation of $\|\mathbf{Z}\|$. (C.6) pertains to the condition in Le Cam's third lemma. It implies that the selection mechanism is asymptotically equivalent to random sampling such that the selection

indicators can be treated as i.i.d. with the law $\prod_{i=1}^n p(R_i; Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)$. Under the original case-cohort design, this condition holds for $p(1; y, \delta, \mathbf{z}, \mathbf{w}) = \delta + (1 - \delta)p_0$, where p_0 is the sampling fraction of the subcohort. Under the nested case-control design, a subject with observation $(y, \delta = 0, \mathbf{z}, \mathbf{w})$ is not selected if and only if he/she is not selected at any observed event time before y , and this probability is $\prod_{Y_i < y} (1 - \Delta_i k/n_i)$, where n_i is the size of the risk set at Y_i , and k is the number of controls selected at each observed event time. Thus, (C.6) holds for $p(0; y, \delta, \mathbf{z}, \mathbf{w}) = (1 - \delta) \exp[-kE\{\Delta/\Pr(Y = y)|_{y=Y}\}]$; see the Supplemental Materials.

We state the asymptotic results in two theorems, which are proved in the Appendix.

Theorem 1—Under conditions (A.1)–(A.2) and (C.1)–(C.7), $|\hat{\theta} - \theta_0| + \sup_{[0, \tau]} |\hat{\Lambda}(t) - \Lambda_0(t)| + \sup_{\mathbf{x}, \mathbf{z}} |F(\mathbf{x}, \mathbf{z}) - F_0(\mathbf{x}, \mathbf{z})| \rightarrow 0$ almost surely.

Theorem 2—Under conditions (A.1)–(A.2) and (C.1)–(C.7), $n^{1/2}(\hat{\theta} - \theta_0)$ converges in distribution to a zero-mean normal random vector whose covariance matrix attains the semiparametric efficiency bound.

As justified at the end of the Appendix, the limiting covariance matrix of $\hat{\theta}$ can be estimated by the negative inverse of the second-order difference of the profile likelihood function for θ . Specifically, we hold θ fixed in the EM algorithm and set the profile likelihood function $p\hat{f}(\theta)$ to the value of $l_n(\theta, \Lambda, p_{sj})$ at the convergence. Then the covariance matrix of $\hat{\theta}$ is estimated by the negative inverse of the matrix whose (k, l) th element is

$$h_n^{-2} \left\{ p\hat{f}(\hat{\theta} + \mathbf{e}_k h_n + \mathbf{e}_l h_n) - p\hat{f}(\hat{\theta} + \mathbf{e}_k h_n) - p\hat{f}(\hat{\theta} + \mathbf{e}_l h_n) + p\hat{f}(\hat{\theta}) \right\},$$

where $\mathbf{e}_k$ is the k th canonical vector, and h_n is a perturbation parameter typically set to $n^{-1/2}$.

3. SIMULATION STUDIES

We conducted simulation studies to assess the performance of the proposed methods in realistic situations. In the first set of studies, we set $X = I(U_1 < 0.3)$, $Z = [100U_2 \sin(U_1)/\sin(1)]$ and $W = U_3$, where U_1 , U_2 and U_3 are independent Uniform(0, 1), and $[x]$ denotes the largest integer no larger than x . The choice of Z , which has 100 values, mimics a continuous covariate, such as age, that is recorded in discrete values in practice. We generated the event times under the proportional hazards model $\Lambda(t|X, Z, W) = 0.1te^{-0.5X+0.5Z+0.8W}$ or the proportional odds model $\Lambda(t|X, Z, W) = \log(1+0.2te^{-0.5X+0.5Z+0.8W})$. The censoring time distribution was set to Uniform[0, 3], yielding approximately 80% censored observations.

We let $n = 10000$ and adopted the original case-cohort sampling (Prentice, 1986) with 0.2 selection probability. In the EM algorithm, we set the initial value of θ to $\mathbf{0}$ and the initial value of $\Lambda\{\cdot\}$ to the inverse of the total number of cases. Since Z is discrete, we estimated $P(x|z)$ by the empirical probability function of X given Z . We estimated the covariance matrix of θ by the profile likelihood method with step size of $6n^{-1/2}$. The results are summarized in Table 1. The parameter estimators are virtually unbiased, the variance estimators accurately reflect the true variations, and the confidence intervals have reasonable coverage probabilities.

The second set of studies was designed to have a continuous covariate that is correlated with a continuous expensive variable. The set-up is the same as the first one except that X is Uniform($-0.5, 0.5$) and $Z = U \sin(X)$, where U is Uniform($0, 1$). We set $n = 2000$ and subcohort selection probability to 0.3. Since Z is continuous, we estimated $P(x|z)$ by the kernel smoothing. We first normalized Z to have mean zero and variance one. We then used the Epanechnikov kernel function, i.e., $K(z, z) = 0.75(1 - \|z - z\|^2)I(\|z - z\| \leq 1)$, and set the bandwidth a_n to $n^{-1/4}$. As shown in Table 1, the proposed methods continue to perform well.

We also evaluated the full-cohort estimator, the Estimator II of Borgan et al. (2000) with time-dependent weights, which is in the spirit of Robins et al. (1994), and the estimator of Zeng et al. (2006), which assumes independence of X and Z . We simulated data in the same manner as in the second set of studies. The results under the proportional hazards model are summarized in Table 2. By comparing the results of Tables 1 and 2, we see that the efficiency loss of the proposed estimator (with subcohort selection probability of 0.3) relative to the full-cohort estimator is fairly small. The relative efficiencies of Borgan et al.'s estimator to the proposed estimator are approximately 90%, 72% and 61% for estimating the effects of X , Z and W , respectively. (The proposed estimator is expected to gain more efficiency for Z and W than for X because it makes full use of the first-phase covariate information.) The bias of Zeng et al.'s estimator is approximately 38% of the true effect size in estimating the effect of X , 74% in estimating the effect of Z and 7% in estimating the effect of W .

We conducted additional simulation studies with smaller n and p_0 and also under a nested case-control design. The results are presented in the Supplemental Materials.

4. REAL EXAMPLES

The National Wilms' Tumor Study Group in the U.S. has conducted a series of studies on Wilms' tumor, a rare kidney cancer in young children. We use data on 3915 subjects from two of the studies (D'Angio et al., 1989; Green et al., 1998) to evaluate the effects of the histological type of the tumor, stage, age at diagnosis, and tumor diameter on relapse-free survival. The censoring rate is approximately 83%.

Each subject's histological type was assessed both by a local pathologist and by an experienced pathologist in a central facility. The latter assessment was more accurate but much more expensive and time-consuming. If a two-phase design had been used, the cost of the studies would have been drastically reduced. To illustrate the benefit of two-phase sampling, we adopt the stratified case-cohort design of Kulich and Lin (2004). Specifically, we divide the subjects into eight strata according to local histology reading (favorable vs unfavorable), stage (III–IV vs I–II), and age (≥ 1 year vs < 1 year); we select all 260 controls in the five smallest strata, 120 controls from the two larger strata, and 160 controls from the largest stratum; all 669 cases are selected.

Following Kulich and Lin (2004), we include a continuous piecewise-linear age effect in the regression analysis. The covariates include the central histology assessment, the linear effect

for $\text{age} < 1$, the linear effect for $\text{age} = 1$ year, the interactions between histology and the two age variables, stage, tumor diameter, and the interaction between stage and tumor diameter. By fitting a logistic regression model to the full data, we find that only stage and tumor diameter are significantly associated with histology. Thus, we let $\mathbf{Z}$ to include stage and tumor diameter, and $\mathbf{W}$ to include all other covariates.

We consider the class of logarithmic transformations. Although tumor diameter is a continuous variable, it only has 26 distinct observed values in the data. With nearly 4000 subjects, we can choose the bandwidth sufficiently small such that the kernel estimation is essentially the estimation of the conditional distribution of histology given each stage and each distinct value of tumor diameter. We draw the subcohort 1000 times. The results of the case-cohort analysis are shown in Table 3, along with the full-cohort analysis, for both the proportional hazards and proportional odds models. The latter model is selected by the AIC criterion, which is tantamount to maximizing the objective function given in (3) because the numbers of parameters for different transformation models are the same.

Even with only one-third of the subjects selected for central histology assessments, the point estimates of the case-cohort analysis are close to their full-cohort counterparts, and the variances are only slightly increased. Kulich and Lin (2004) reported the results of the case-cohort analysis under the proportional hazards model using their weighted method and Borgan et al.'s (2000); Breslow et al. (2009) applied their calibration method to the same data. The standard error estimates of our case-cohort analysis are lower than the other three methods. The case-cohort analysis of Kulich and Lin (2004) and Breslow et al. (2009) used the local histology assessments and thus used more data than our analysis.

For further illustration, we consider the ARIC study (The ARIC Investigators, 1989), which recently adopted a stratified case-cohort design; see the Supplemental Materials.

5. DISCUSSION

We have developed efficient estimators for semiparametric transformation models under general two-phase designs allowing the first-phase covariates to be continuous and correlated with expensive covariates. The likelihood function of interest is not tractable because it involves the conditional density function of expensive covariates given continuous first-phase covariates. We approximated the conditional density function by kernel smoothing. The approximations are different between the subjects with and without measurements of expensive covariates. Because the approximate likelihood function involves nonparametric smoothing estimators, which do not possess the usual parametric convergence rate, the existing nonparametric maximum likelihood estimation (NPMLE) theory (Murphy, 1994; Zeng and Lin, 2010), which utilizes the score operators for both parametric and nonparametric components and treats them equally in the inference, is not applicable. Through innovative use of modern empirical process theory and kernel estimation theory, we were able to establish the consistency, weak convergence and asymptotic efficiency of the proposed estimators with mixed convergence rates. Our proofs start with the efficient score function for the regression parameters and chooses appropriate bandwidths to control the asymptotic bias due to kernel smoothing.

It is not feasible to maximize the kernel-smoothed log-likelihood function directly or by applying the existing EM algorithms for NPMLE methods. We developed a novel semiparametric EM algorithm by artificially creating a latent variable for non-selected subjects such that the high-dimensional cumulative baseline hazard function and conditional distribution function can be evaluated explicitly in the M-step of the EM algorithm; the only major calculation is to solve a small number of equations for the regression parameters. We

established the ascent property of the proposed EM algorithm to ensure its good performance.

In most two-phase studies, such as the ARIC study, the sample sizes are very large and there are only a small number of distinct observed values for $\mathbf{Z}$. Then we can treat $\mathbf{Z}$ as discrete random variables and choose the bandwidth sufficiently small so that the kernel estimation is essentially equivalent to the use of empirical probability functions. Our theory still holds, but the computation is faster.

Due to curse of dimensionality, $\mathbf{Z}$ cannot be very high-dimensional, i.e., > 3 . There are several ways to obtain a low-dimensional $\mathbf{Z}$. First, if there is prior scientific knowledge about the dependence among the covariates, then that information can be incorporated into the modeling. Second, we may assess the dependence empirically from the data. Third, we may adopt a dimension-reduction technique, such as principal components analysis.

The bandwidth a_n in (3) can be chosen in a data-adaptive manner. Specifically, we can choose a_n through cross-validation. For any fixed a_n , we use part of the data as the test data and the remainder as the validation data. We evaluate (3) in the validation data using the estimates obtained from the test data. The optimal bandwidth a_n is the value that maximizes the average cross-validation likelihood.

We have assumed that the covariates are time-independent. Our results can be readily modified if there are external time-dependent covariates in $\mathbf{W}$. Generalization to the case of time-dependent $\mathbf{Z}$ is challenging due to the difficulties in estimating the conditional density of $\mathbf{X}$ given time-dependent covariates. One possible strategy is to jointly model the event time T and the latent trajectory of $\mathbf{Z}$ over time while allowing $\mathbf{X}$ to depend on the trajectory of $\mathbf{Z}$.

Assumption (A.2) allows the missingness on $\mathbf{X}$ to depend only on the first-phase data $\mathbf{O} \equiv (Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)$ ($i = 1, \dots, n$). This assumption holds naturally for all two-phase studies. However, the proposed likelihood function and the corresponding inference procedures remain valid under the general missing-at-random framework, which allows missingness to depend on all observed data, including observed $\mathbf{X}$.

Under (A.1), C is allowed to depend on observed $\mathbf{X}$ but not on unobserved $\mathbf{X}$. If this coarsening-at-random assumption fails, we specify that the cumulative hazard function of C satisfies a semiparametric transformation model:

$\Lambda_c(t|\mathbf{X}, \mathbf{Z}, \mathbf{W}) = G_c\{\Lambda_c(t)e^{\beta_c^T \mathbf{X} + \gamma_c^T \mathbf{Z} + \eta_c^T \mathbf{W}}\}$, where $G_c(\cdot)$ is a known increasing function. Then $P(Y, \Delta|\mathbf{X}, \mathbf{Z}, \mathbf{W})$ in (3) will become the product of expression (2) and the following expression

$$\left[\Lambda'_c(Y) e^{\beta_c^T \mathbf{X} + \gamma_c^T \mathbf{Z} + \eta_c^T \mathbf{W}} G'_c \left\{ \Lambda_c(Y) e^{\beta_c^T \mathbf{X} + \gamma_c^T \mathbf{Z} + \eta_c^T \mathbf{W}} \right\} \right]^{1-\Delta} \exp \left[-G_c \left\{ \Lambda_c(Y) e^{\beta_c^T \mathbf{X} + \gamma_c^T \mathbf{Z} + \eta_c^T \mathbf{W}} \right\} \right].$$

The proposed EM algorithm can still be used but the M-step will now involve estimating the parameters in the censoring distribution.

Our work is the first attempt at providing efficient estimation for two-phase studies with continuous first-phase covariates that are correlated with expensive covariates. Although we focused on cohort studies with censored survival data, our approach can be applied to any two-phase studies, particularly those with dichotomous and continuous outcomes. We simply replace $P(Y, \Delta|\mathbf{X}, \mathbf{Z}, \mathbf{W})$ in (3) by the conditional density function of the new

outcome variable given covariates. Our approach can also be extended to multivariate failure time data (Lu and Shih, 2006; Kang and Cai, 2009; Zhang et al., 2011).

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by the National Institutes of Health awards R37GM047845, R01AI029168 and R01CA082659. The authors thank an Associate Editor and three referees for their helpful comments.

References

- Barlow WE. Robust variance estimation for the case-cohort design. *Biometrics*. 1994; 50:1064–1072. [PubMed: 7786988]
- Borgan Ø, Langholz B, Samuelsen SO, Goldstein L, Pagoda J. Exposure stratified case-cohort designs. *Lifetime Data Analysis*. 2000; 6:39–58. [PubMed: 10763560]
- Breslow NE, Lumley T, Ballantyne CM, Chambless LE, Kulich M. Using the whole cohort in the analysis of case-cohort data. *American Journal of Epidemiology*. 2009; 169:1398–1405. [PubMed: 19357328]
- Cai T, Zheng Y. Non-parametric evaluation of biomarker accuracy under nested case-control studies. *Journal of the American Statistical Association*. 2011; 106:569–580. [PubMed: 22844169]
- Cai T, Zheng Y. Evaluating Prognostic Accuracy of biomarkers under Nested Case-control Studies. *Biostatistics*. 2012; 13:89–100. [PubMed: 21856652]
- Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 1999; 94:896–908.
- Chen K. Generalized case-cohort sampling. *Journal of the Royal Statistical Society, Series B*. 2001; 63:791–809.
- Chen K, Jin Z, Ying Z. Semiparametric analysis of transformation models with censored data. *Biometrika*. 2002; 89:659–668.
- Chen K, Lo SH. Case-cohort and case-control analysis with Cox's model. *Biometrika*. 1999; 86:755–764.
- Cox DR. Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*. 1972; 34:187–220.
- Cox DR. Partial Likelihood. *Biometrika*. 1975; 62:269–276.
- D'Angio GJ, Breslow N, Bechwith JB, Evans A, Baum E, deLorimier A, Fernbach D, Hrabovsky E, Jones B, Kelalis P, Otherson HB, Tefft M. Treatment of Wilms' tumor. Results of the third national Wilms' tumor study. *Cancer*. 1989; 64:349–360. [PubMed: 2544249]
- Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. *The Annals of Statistics*. 1992; 20:69–79.
- Green DM, Breslow NE, Bechwith JB, Finklestein JZ, Grundy PE, Thomas PR, Kim T, Shochat SJ, Haase GM, Ritchey ML, Kelalis PP, D'Angio GJ. Comparison between single-dose and divided-dose administration of dactinomycin and doxorubicin for patients with Wilms' tumor: a report from the national Wilms' Tumor Study Group. *Journal of Clinical Oncology*. 1998; 16:237–245. [PubMed: 9440748]
- Johansen S. An extension of Cox's regression model. *International Statistical Review*. 1983; 51:258–262.
- Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*. 1988; 7:149–160. [PubMed: 3353602]
- Kang S, Cai J. Marginal hazards model for case-cohort studies with multiple disease outcomes. *Biometrika*. 2009; 96:887–901. [PubMed: 23946547]
- Kong L, Cai J. Case-cohort analysis with accelerated failure time model. *Biometrics*. 2009; 65:135–142. [PubMed: 18537948]

- Kong L, Cai J, Sen PK. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*. 2004; 91:305–319.
- Kulich M, Lin DY. Additive hazards regression for case-cohort studies. *Biometrika*. 2000; 87:73–87.
- Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association*. 2004; 99:832–844.
- Langholz B, Borgan Ø. Counter-matching: A stratified nested case-control sampling method. *Biometrika*. 1995; 82:69–79.
- Lu SE, Shih JH. Case-cohort designs and analysis for clustered failure time data. *Biometrics*. 2006; 62:1138–1148. [PubMed: 17156289]
- Lu W, Tsiatis AA. Semiparametric transformation models for the case-cohort study. *Biometrika*. 2006; 93:207–214.
- Luo X, Tsai WY, Xu Q. Pseudo-partial likelihood estimators for the Cox regression model with missing covariates. *Biometrika*. 2009; 96:617–633. [PubMed: 23946546]
- Martinussen T. Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics*. 1999; 26:479–491.
- Murphy SA. Consistency in a proportional hazards model incorporating a random effect. *The Annals of Statistics*. 1994; 22:712–731.
- Murphy SA, van der Vaart AW. On the profile likelihood. *Journal of the American Statistical Association*. 2000; 95:449–465.
- Nan B. Efficient estimation for case-cohort studies. *Canadian Journal of Statistics*. 2004; 32:403–419.
- Nan B, Yu M, Kalbfleisch JD. Censored linear regression for case-cohort studies. *Biometrika*. 2006; 93:747–762.
- Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Qi L, Wang CY, Prentice RL. Weighted estimators for proportional hazards regression with missing covariates. *J Am Statist Assoc*. 2005; 100:1250–1263.
- Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
- Rudin, W. *Functional Analysis*. New York: McGraw-Hill; 1973.
- Scheike TH, Juul A. Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics*. 2004; 5:193–206. [PubMed: 15054025]
- Scheike TH, Martinussen T. Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavian Journal of Statistics*. 2004; 31:283–293.
- Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. *The Annals of Statistics*. 1988; 16:64–81.
- The ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *American Journal of Epidemiology*. 1989; 129:687–702. [PubMed: 2646917]
- Thomas, D. Addendum to: Methods of cohort analysis: appraisal by application to asbestos mining. In: Liddell, FDK.; McDonald, JC.; Thomas, DC., editors. *Journal of the Royal Statistical Society, Series A*. Vol. 140. 1977. p. 469–491.
- van der Vaart, AW.; Wellner, JA. *Weak Convergence and Empirical Processes*. New York: Springer-Verlag; 1996.
- Zeng D, Lin DY, Avery CL, North KE, Bray MS. Efficient semipara-metric estimation of haplotype-disease associations in case-cohort and nested case-control studies. *Biostatistics*. 2006; 7:486–502. [PubMed: 16500923]
- Zeng D, Lin DY. Maximum likelihood estimation in semiparametric regression models with censored data (with discussion). *Journal of the Royal Statistical Society, Series B*. 2007; 69:507–564.
- Zeng D, Lin DY. A general asymptotic theory for maximum likelihood estimation in semiparametric regression models with censored data. *Statistica Sinica*. 2010; 20:871–910. [PubMed: 20577580]
- Zhang H, Schaubel DE, Kalbfleisch JD. Proportional hazards regression for the analysis of clustered survival data from case-cohort studies. *Biometrics*. 2011; 67:18–29. [PubMed: 20560939]

APPENDIX: PROOFS OF THEOREMS

Before proving Theorems 1 and 2, we show that the conclusions of the theorems hold under the sampling mechanism satisfying condition (C.6) if they hold under independent sampling. Let $\mathcal{P}_n$ denote the probability measure

$\prod_{i=1}^n P(Y_i, \Delta_i, R_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i) \prod_{i=1}^n p(R_i | Y_i, \Delta_i, \mathbf{Z}_i, \mathbf{W}_i)$, and let $\mathcal{Q}_n$ denote the probability measure associated with the observed data. It follows from condition (C.6) that under $\mathcal{P}_n$, the likelihood ratio statistic between the joint density of the observed data and the density given by $\mathcal{P}_n$, i.e., $d\mathcal{Q}_n/d\mathcal{P}_n$, converges in distribution to some random variable $\mathcal{Z}$, whose conditional distribution given $\mathbf{O}$ is approximated by $e^{-\mathcal{A}}$, where $\mathcal{A}$ is normal with mean $-f(\mathbf{O})/2$ and variance $f(\mathbf{O})$. Suppose that $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to some random variable $\mathcal{L}$ under $\mathcal{P}_n$. Then for any subsequence $\{n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0), d\mathcal{Q}_n/d\mathcal{P}_n\}$, the tightness implies that there exists a subsequence which converges to $(\mathcal{L}, \mathcal{Z})$. Therefore, Le Cam's third lemma entails that, under probability measure $\mathcal{Q}_n$, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to a probability measure given by $E\{I(\mathcal{L} \in B) | \mathcal{Z}\}$ for any Borel set B . Because $E(\mathcal{Z} | \mathbf{O}) = 1$ and $\boldsymbol{\theta}_0$ only depends on $\mathbf{O}$, we have $E\{I(\mathcal{L} \in B) | \mathcal{Z}\} = E\{I(\mathcal{L} \in B)\}$. Thus, under the measure actually generating the data, $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ converges in distribution to the same random variable $\mathcal{L}$. By the sub-subsequence argument, this convergence holds for the whole sequence of $n^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$. Thus, our claim is established. Because of this result, we assume in the following proofs that the observations $(Y_i, \Delta_i, R_i | \mathbf{X}_i, \mathbf{Z}_i, \mathbf{W}_i)$ ($i = 1, \dots, n$) are i.i.d.

Proof of Theorem 1

We first show that $\limsup_n \Lambda(\hat{v}) < \infty$ with probability one. We define a distribution function $P_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$ with point mass q_{is}/n at the observation $(Y_i, \Delta_i, \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i)$ ($i = 1, \dots, n$; $s = 1, \dots, m$), where $q_{is} = I(\mathbf{X}_i = \mathbf{x}_s)$ for $i \in \mathcal{C}$ and

$$\hat{q}_{is} = \frac{P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i) \sum_{j=1}^n w_{ji} p_{sj}}{\sum_{s=1}^m P(Y_i, \Delta_i | \mathbf{X} = \mathbf{x}_s, \mathbf{Z}_i, \mathbf{W}_i) \sum_{j=1}^n w_{ji} p_{sj}}$$

for $i \in \bar{\mathcal{C}}$. The M-step of the EM algorithm implies that $(\hat{\boldsymbol{\theta}}, \hat{\Lambda})$ maximizes

$$\int \log P(Y, \Delta | \mathbf{X}, \mathbf{Z}, \mathbf{W}) d\hat{P}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}). \quad (\text{A.1})$$

Differentiating (A.1) with respect to the jump size of Λ yields

$$\hat{\Lambda}(t) = -n^{-1} \sum_{Y_i \leq t} \frac{\Delta_i}{\int I(Y \geq Y_i) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) d\hat{P}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})},$$

where

$$\mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) = e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \left[\Delta \frac{G'' \left\{ \Lambda(Y) e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \right\}}{G' \left\{ \Lambda(Y) e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \right\}} - G' \left\{ \Lambda(Y) e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \right\} \right].$$

Define

$$\tilde{\Lambda}(t) = -n^{-1} \sum_{Y_i \leq t} \frac{\Delta_i}{\int I(Y \geq Y_i) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) dP_{10}(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})},$$

where $P_{10}(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$ is the distribution function for $(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$. We now show that $\tilde{\Lambda}(t)$ converges uniformly to $\Lambda_0(t)$. The classes of functions $\{I(Y \geq t) : t \in [0, \tau]\}$ and

$$\{\int I(Y \geq t) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) dP_{10}(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}) : t \in [0, \tau]\}$$

are P-Donsker. After integration by part, the denominator of $\tilde{\Lambda}(t)$ can be expressed as

$$A(Y_i) = -E \left[G' \left\{ \Lambda_0(t) e^{\beta_0^T \mathbf{X} + \gamma_0^T \mathbf{Z} + \eta_0^T \mathbf{W}} \right\} e^{\beta_0^T \mathbf{X} + \gamma_0^T \mathbf{Z} + \eta_0^T \mathbf{W}} e^{-G \left\{ \Lambda_0(t) e^{\beta_0^T \mathbf{X} + \gamma_0^T \mathbf{Z} + \eta_0^T \mathbf{W}} \right\}} S_c(t | \mathbf{X}, \mathbf{Z}, \mathbf{W}) \right] \Big|_{t=Y_i},$$

where $S_c(t | \mathbf{X}, \mathbf{Z}, \mathbf{W})$ is the conditional survival function of the censoring time given covariates. Condition (C.4) implies that $A(\cdot)$ is bounded from below by a positive constant. Thus, the Glivenko-Cantelli theorem entails that

$$\sup_{t \in [0, \tau]} \left| \tilde{\Lambda}(t) + E \left\{ \frac{\Delta I(Y \leq t)}{A(Y)} \right\} \right| \rightarrow 0$$

with probability one. On the other hand, $E \{ \Delta I(Y \leq t) / A(Y) \} = -\Lambda_0(t)$. Hence, $\tilde{\Lambda}$ converges to Λ_0 uniformly in $[0, \tau]$ with probability one.

By the EM algorithm,

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \sum_{s=1}^m \hat{q}_{is} \left[\Delta_i \log \hat{\Lambda}\{Y_i\} + \Delta_i \log G'(\hat{\Lambda}(Y_i) e^{\hat{\beta}^T \mathbf{x}_s + \hat{\gamma}^T \mathbf{Z}_i + \hat{\eta}^T \mathbf{W}_i}) - G(\hat{\Lambda}(Y_i) e^{\hat{\beta}^T \mathbf{x}_s + \hat{\gamma}^T \mathbf{Z}_i + \hat{\eta}^T \mathbf{W}_i}) \right] \\ & \geq n^{-1} \sum_{i=1}^n \sum_{s=1}^m \hat{q}_{is} \left[\Delta_i \log \tilde{\Lambda}\{Y_i\} + \Delta_i \log G'(\tilde{\Lambda}(Y_i) e^{\beta_0^T \mathbf{x}_s + \gamma_0^T \mathbf{Z}_i + \eta_0^T \mathbf{W}_i}) - G(\tilde{\Lambda}(Y_i) e^{\beta_0^T \mathbf{x}_s + \gamma_0^T \mathbf{Z}_i + \eta_0^T \mathbf{W}_i}) \right]. \end{aligned}$$

The construction of $\tilde{\Lambda}$ implies that $\tilde{\Lambda}\{Y_i\} \geq c_0/n$ when $\Delta_i = 1$ for some positive constant c_0 .

Since $\tilde{\Lambda}$ uniformly converges to Λ_0 in $[0, \tau]$ and $\sum_{s=1}^m \hat{q}_{is} = 1$, the right-hand side of the above inequality is equal to $-\log n + O_p(1)$. In light of condition (C.3) that $(1+G'(x)) \exp\{-G(x)\} \geq c_1(1+x)^{-\nu_0}$ for some constant $\nu_0 > 0$ and $c_1 > 0$, we have

$$n^{-1} \sum_{i=1}^n \sum_{s=1}^m \hat{q}_{is} \left[\Delta_i \log \left(n \hat{\Lambda}\{Y_i\} \right) - \nu_0 \log \left\{ 1 + \hat{\Lambda}(Y_i) e^{\hat{\beta}^T \mathbf{x}_s + \hat{\gamma}^T \mathbf{Z}_i + \hat{\eta}^T \mathbf{W}_i} \right\} \right] \geq O_p(1).$$

By the boundedness of $(\hat{\beta}^T \mathbf{x}_s + \hat{\gamma}^T \mathbf{Z}_i + \hat{\eta}^T \mathbf{W}_i)$ and the constraint that $\sum_{s=1}^m \hat{q}_{is} = 1$, we have

$$n^{-1} \sum_{i=1}^n \left[\Delta_i \log \left(n \hat{\Lambda} \{Y_i\} \right) - \nu_0 \log \{c_2 + \hat{\Lambda}(Y_i)\} \right] \geq O(1)$$

for some positive constant c_2 . Thus, we obtain the same inequality as (6.3) of Zeng and Lin (2010). Using the partitioning technique of Murphy (1994) and Zeng and Lin (2010), we can then show that the left-hand side of the above inequality will diverge to $-\infty$, which is a contradiction, unless $\limsup \hat{\Lambda}(t) < \infty$.

We choose a convergent subsequence from the sequence of estimators $(\hat{\theta}, \hat{\Lambda})$ implicitly indexed by n . Since Λ is non-increasing and bounded, Helly's selection theorem allows us to choose a further subsequence from any subsequence of the estimators, still denoted as $(\hat{\theta}, \hat{\Lambda})$, such that $\hat{\theta} \rightarrow \theta^*$ and $\hat{\Lambda}$ converges to Λ^* uniformly for some vector θ^* and some function Λ^* . Furthermore, we can choose a subsequence $F(\mathbf{x}, \mathbf{z})$ such that $F(\mathbf{x}, \mathbf{z})$ converges to some function $F^*(\mathbf{x}, \mathbf{z})$ in the norm of the bounded variation space. Likewise, the measure P_1 converges weakly to some finite measure. The consistency will hold if we can show that $\theta^* = \theta_0$, $\Lambda^* = \Lambda_0$ and $F^* = F_0$.

To that end, we establish some facts about F and F^* . Since

$$\begin{aligned} \left| \hat{F}(\mathbf{x}, \mathbf{z}_1) - \hat{F}(\mathbf{x}, \mathbf{z}_2) \right| &= \left| n^{-1} \sum_{j=1}^n \sum_{i \in \mathcal{C}} \hat{p}_{ij} I(\mathbf{X}_i \leq \mathbf{x}) \{I(\mathbf{Z}_i \leq \mathbf{z}_2) - I(\mathbf{Z}_i \leq \mathbf{z}_1)\} \right| \\ &\leq n^{-1} \sum_{j=1}^n |I(\mathbf{Z}_i \leq \mathbf{z}_2) - I(\mathbf{Z}_i \leq \mathbf{z}_1)|, \end{aligned}$$

we conclude that $\sup_{\mathbf{x}} |F^*(\mathbf{x}, \mathbf{z}_1) - F^*(\mathbf{x}, \mathbf{z}_2)| \leq E[|I(\mathbf{Z}_i \leq \mathbf{z}_2) - I(\mathbf{Z}_i \leq \mathbf{z}_1)|]$, so $F^*(\mathbf{x}, \mathbf{z})$ is uniformly continuous in $\mathbf{z}$. Since $\hat{p}_{sj}$ maximizes (3), differentiating (3) with respect to $\hat{p}_{sj}$ yields

$$\frac{\sum_{i \in \mathcal{C}} I(\mathbf{X}_i = \mathbf{x}_s) w_{ji}}{\hat{p}_{sj}} + \sum_{i \in \mathcal{C}} \frac{\mathcal{B}_2(Y_i, \Delta_i, \mathbf{x}_s, \mathbf{Z}_j, \mathbf{W}_i; \hat{\theta}, \hat{\Lambda}) w_{ji}}{\sum_{k=1}^n \sum_{s=1}^m \mathcal{B}_2(Y_i, \Delta_i, \mathbf{x}_s, \mathbf{Z}_k, \mathbf{W}_i; \hat{\theta}, \hat{\Lambda}) w_{ki} \hat{p}_{sk}} - \hat{\mu}_j = 0, \quad (\text{A. } 2)$$

where

$$\mathcal{B}_2(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta, \Lambda) = \left[e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \theta^T \mathbf{W}} G' \left\{ \Lambda(Y) e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \right\} \right]^\Delta \exp \left\{ -G \left(e^{\beta^T \mathbf{X} + \gamma^T \mathbf{Z} + \eta^T \mathbf{W}} \right) \right\}$$

and $\hat{\mu}_j$ is the Lagrange multiplier for the constraint that $\sum_{s=1}^m \hat{p}_{sj} = 1$. Thus,

$$\hat{p}_{sj} = \sum_{i \in \mathcal{C}} I(\mathbf{X}_i = \mathbf{x}_s) w_{ji} / \left\{ \hat{\mu}_j - \sum_{i \in \mathcal{C}} \frac{\mathcal{B}_2(Y_i, \Delta_i, \mathbf{x}_s, \mathbf{Z}_j, \mathbf{W}_i; \hat{\theta}, \hat{\Lambda}) w_{ji}}{\sum_{k=1}^n \sum_{s=1}^m \mathcal{B}_2(Y_i, \Delta_i, \mathbf{x}_s, \mathbf{Z}_k, \mathbf{W}_i; \hat{\theta}, \hat{\Lambda}) w_{ki} \hat{p}_{sk}} \right\}. \quad (\text{A. } 3)$$

Consequently,

$$\hat{F}(\mathbf{x}, \mathbf{z}) = n^{-1} \sum_{j=1}^n \sum_{i \in \mathcal{C}} \frac{I(\mathbf{Z}_j \leq \mathbf{z}, \mathbf{X}_i \leq \mathbf{x}) w_{ji}}{\mathcal{B}_{3n}(\mathbf{X}_i, \mathbf{Z}_j; \hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F})},$$

where

$$\mathcal{B}_{3n}(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) = \hat{\mu}(\mathbf{z}) - n^{-1} \sum_{k \in \mathcal{C}} \frac{\mathcal{B}_2(Y_k, \Delta_k, \mathbf{x}, \mathbf{z}, \mathbf{W}_k; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\mathbf{z} - \mathbf{Z}_k)/a_n\}}{\int \mathcal{B}_2(Y_k, \Delta_k, \tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \mathbf{W}_k; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\tilde{\mathbf{z}} - \mathbf{Z}_k)/a_n\} d\hat{F}(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})},$$

and $\mu(\hat{\mathbf{z}})$ is a function satisfying

$$\begin{aligned} & n^{-1} \sum_{i \in \mathcal{C}} \frac{K\{(\mathbf{z} - \mathbf{Z}_i)/a_n\}}{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z}_i)/a_n\}} \\ & \times \left[\mu(\mathbf{z}) - n^{-1} \sum_{k \in \mathcal{C}} \frac{\mathcal{B}_2(Y_k, \Delta_k, \mathbf{x}, \mathbf{z}, \mathbf{W}_k; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\mathbf{z} - \mathbf{Z}_k)/a_n\}}{\int \mathcal{B}_2(Y_k, \Delta_k, \tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \mathbf{W}_k; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\tilde{\mathbf{z}} - \mathbf{Z}_k)/a_n\} d\hat{F}(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})} \right]^{-1} - 1 = 0 \quad (\text{A.4}) \end{aligned}$$

with $\mathcal{B}_{3n}(\mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) > 0$ and $\mu(\hat{\mathbf{Z}}_j) = \mu_j$. Note that

$$a_n^{-d_{\mathbf{z}}} \int_{\mathbf{x}, \mathbf{z}} \mathcal{B}_2(y, \delta, \mathbf{x}, \mathbf{z}, \mathbf{w}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\mathbf{z} - \mathbf{z}_0)/a_n\} d\hat{F}(\mathbf{x}, \mathbf{z}) = \int_{\mathbf{x}, \tilde{\mathbf{z}}} \mathcal{B}_2(y, \delta, \mathbf{x}, \mathbf{z}_0 + a_n \tilde{\mathbf{z}}, \mathbf{w}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K(\tilde{\mathbf{z}}) d\hat{F}(\mathbf{x}, \mathbf{z}_0 + a_n \tilde{\mathbf{z}}).$$

Thus,

$$\begin{aligned} & \sup_{y, \delta, \mathbf{w}, \mathbf{z}_0} \left| a_n^{-d_{\mathbf{z}}} \int_{\mathbf{x}, \mathbf{z}} \mathcal{B}_2(y, \delta, \mathbf{x}, \mathbf{z}, \mathbf{w}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\mathbf{z} - \mathbf{z}_0)/a_n\} d\hat{F}(\mathbf{x}, \mathbf{z}) \right. \\ & \quad - \int_{\mathbf{x}, \tilde{\mathbf{z}}} \mathcal{B}_2(y, \delta, \mathbf{x}, \mathbf{z}_0 \\ & \quad + a_n \tilde{\mathbf{z}}, \mathbf{w}; \boldsymbol{\theta}^*, \Lambda^*) K(\tilde{\mathbf{z}}) dF^*(\mathbf{x}, \mathbf{z}_0 \\ & \quad \left. + a_n \tilde{\mathbf{z}}) \right| \rightarrow 0. \end{aligned}$$

Using integration by part and the uniform continuity of $F^*(\mathbf{x}, \mathbf{z})$ in $\mathbf{z}$, we can show that the second term on the left-hand side has a uniform limit, denoted by $B^*(y, \delta, \mathbf{w}, \mathbf{z}_0)$. Therefore,

$$n^{-1} \sum_{k \in \mathcal{C}} \frac{\mathcal{B}_2(Y_k, \Delta_k, \mathbf{X}_i, \mathbf{z}, \mathbf{W}_k; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\mathbf{z} - \mathbf{Z}_k)/a_n\}}{\int \mathcal{B}_2(Y_k, \Delta_k, \tilde{\mathbf{x}}, \tilde{\mathbf{z}}, \mathbf{W}_k; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) K\{(\tilde{\mathbf{z}} - \mathbf{Z}_k)/a_n\} d\hat{F}(\tilde{\mathbf{x}}, \tilde{\mathbf{z}})}$$

is asymptotically equivalent to

$$n^{-1} \sum_{k \in \mathcal{C}} \frac{\mathcal{B}_2(Y_k, \Delta_k, \mathbf{X}_i, \mathbf{z}, \mathbf{W}_k; \boldsymbol{\theta}^*, \Lambda^*) a_n^{-d_{\mathbf{z}}} K\{(\mathbf{z} - \mathbf{Z}_k)/a_n\}}{B^*(Y_k, \Delta_k, \mathbf{W}_k, \mathbf{Z}_k)},$$

and this equivalence holds uniformly in $\mathbf{X}_i$ and $\mathbf{z}$. Consequently, the left-hand side of equation (A.4) converges uniformly in $\mathbf{z}$ to

$$E \left[\frac{R}{\mu(\mathbf{z}) - E \{ (1-R) \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta^*, \Lambda^*) / B^*(Y, \Delta, \mathbf{W}, \mathbf{z}) | \mathbf{Z} = \mathbf{z} \} } \Big| \mathbf{Z} = \mathbf{z} \right] - 1 = 0.$$

This equation is monotone in $\mu(\mathbf{z})$ and has a unique solution, denoted by $\mu^*(\mathbf{z})$, satisfying that $\mu(\mathbf{z}) - E \{ (1-R) \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta^*, \Lambda^*) / B^*(Y, \Delta, \mathbf{W}, \mathbf{z}) | \mathbf{Z} = \mathbf{z} \} |_{\mathbf{x}=\mathbf{X}}$ is bounded away from zero. By the inverse mapping theorem, the solution to (A.4), $\mu(\mathbf{z})$, converges to $\mu^*(\mathbf{z})$ uniformly in $\mathbf{z}$. In addition, $\mathcal{B}_{3n}(\mathbf{x}, \mathbf{z}; \hat{\theta}, \hat{\Lambda}, \hat{F})$ converges to

$$\mu^*(\mathbf{z}) - E \{ (1-R) \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta^*, \Lambda^*) / B^*(Y, \Delta, \mathbf{W}, \mathbf{z}) | \mathbf{Z} = \mathbf{z} \}$$

for $(\mathbf{x}, \mathbf{z})$ in the support of $(\mathbf{X}, \mathbf{Z})$ and thus has a strictly positive lower bound when n is large enough.

Our final step is to prove that $\theta^* = \theta_0$, $\Lambda^* = \Lambda_0$ and $F^* = F_0$ through the Kullback-Leibler inequality. Let $p_{sj} = \sum_{i \in \mathcal{C}} I(\mathbf{X}_i = \mathbf{x}_s) w_{sj} / \sum_{i \in \mathcal{C}} w_{ij}$, and $\tilde{F}(\mathbf{x}, \mathbf{z}) = n^{-1} \sum_{s=1}^m \sum_{j=1}^n I(\mathbf{x}_s \leq \mathbf{x}, \mathbf{Z}_j \leq \mathbf{z}) \tilde{p}_{sj}$. By the theory on kernel density estimation, $F(\mathbf{x}, \mathbf{z})$ uniformly converges to $F_0(\mathbf{x}, \mathbf{z})$. Furthermore, it follows from the above facts on $F(\mathbf{x}, \mathbf{z})$ that $F(\mathbf{x}, \mathbf{z})$ is absolutely continuous with respect to, $F(\mathbf{x}, \mathbf{z})$, and

$$\hat{F}(\mathbf{x}, \mathbf{z}) = \int \frac{\sum_{i \in \mathcal{C}} K\{(\mathbf{Z}_i - \mathbf{z})/a_n\} / \sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{z})/a_n\}}{\mathcal{B}_{3n}(\mathbf{x}, \mathbf{z}; \hat{\theta}, \hat{\Lambda}, \hat{F})} d\tilde{F}(\mathbf{x}, \mathbf{z}).$$

Thus, $d\hat{F}(\mathbf{x}, \mathbf{z})/d\tilde{F}(\mathbf{x}, \mathbf{z})$ converges uniformly to $dF^*(\mathbf{x}, \mathbf{z})/dF_0(\mathbf{x}, \mathbf{z})$, and $F^*(\mathbf{x}, \mathbf{z})$ is differentiable with respect to $\mathbf{x}$ and $\mathbf{z}$. By definition, $n^{-1} l_n(\hat{\theta}, \hat{\Lambda}, p_{sj}) \rightarrow n^{-1} l_n(\theta_0, \Lambda_0, p_{sj})$. That is,

$$\begin{aligned} & n^{-1} \sum_{i=1}^n \log \frac{\hat{\Lambda}\{Y_i\}}{\Lambda\{Y_i\}} + n^{-1} \sum_{i \in \mathcal{C}} \log \frac{\mathcal{B}_2(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{W}_i; \hat{\theta}, \hat{\Lambda})}{\mathcal{B}_2(Y_i, \Delta_i, \mathbf{X}_i, \mathbf{W}_i; \theta_0, \Lambda_0)} \\ & + n^{-1} \sum_{i \in \mathcal{C}} \log \frac{\int \mathcal{B}_2(Y_i, \Delta_i, \mathbf{x}, \mathbf{z}, \mathbf{W}_i; \hat{\theta}, \hat{\Lambda}) K((\mathbf{z} - \mathbf{Z}_i)/a_n) d\hat{F}(\mathbf{x}, \mathbf{z})}{\int \mathcal{B}_2(Y_i, \Delta_i, \mathbf{x}, \mathbf{z}, \mathbf{W}_i; \theta_0, \Lambda_0) K((\mathbf{z} - \mathbf{Z}_i)/a_n) d\tilde{F}(\mathbf{x}, \mathbf{z})} \\ & + n^{-1} \sum_{i \in \mathcal{C}} \sum_{j=1}^n w_{ji} \log \frac{\hat{p}_{ij}}{\tilde{p}_{ij}} \geq 0. \end{aligned} \quad (\text{A.5})$$

For the first term in (A.5), we note that

$$\frac{\hat{\Lambda}\{Y_i\}}{\Lambda\{Y_i\}} = \frac{\int I(Y \geq Y_i) \mathcal{B}_1(\mathbf{O}; \theta_0, \Lambda_0) dP_{10}}{\int I(Y \geq Y_i) \mathcal{B}_1(\mathbf{O}; \hat{\theta}, \hat{\Lambda}) d\hat{P}_1},$$

which converges uniformly to some positive function. This result implies that Λ^* is absolutely continuous with respect to Λ_0 and the limit of the above expression is equal to

$d\Lambda^*/d\Lambda_0$ at Y_i . Thus, the first term of (A.4) converges almost surely to $E[\Lambda^{*'}(Y)/\Lambda_0'(Y)]$. The convergence of $\tilde{\Lambda}$ and Λ implies that the limit of the second term in (A.5) is

$$E \left[R \log \frac{\mathcal{B}_2(Y, \Delta, \mathbf{X}, \mathbf{W}; \boldsymbol{\theta}^*, \Lambda^*)}{\mathcal{B}_2(Y, \Delta, \mathbf{X}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0)} \right].$$

Through integration by part and kernel approximation, the third term in (A.5) converges to

$$E \left[(1-R) \log \frac{\int_{\mathbf{x}} \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}^*, \Lambda^*) F^*(d\mathbf{x}, \mathbf{z})}{\int_{\mathbf{x}} \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda) F(d\mathbf{x}, \mathbf{z})} \middle| \mathbf{Z}=\mathbf{z} \right].$$

By (A.3), $p_{sj}/p_{sj}^* = dF(\hat{\mathbf{x}}, \mathbf{z})/dF(\mathbf{x}, \mathbf{z})$ at $\mathbf{x} = \mathbf{x}_s$ and $\mathbf{z} = \mathbf{Z}_j$. It follows that p_{sj}/p_{sj}^* converges uniformly to $dF^*(\mathbf{x}, \mathbf{z})/dF(\mathbf{x}, \mathbf{z})|_{\mathbf{x}=\mathbf{x}_s, \mathbf{z}=\mathbf{Z}_j}$. As a result, the last term of (A.5) converges to

$$E \left[R \log \frac{dF^*(\mathbf{X}, \mathbf{z})}{dF_0(\mathbf{X}, \mathbf{z})} \middle| \mathbf{Z}=\mathbf{z} \right].$$

Combining the above results for (A.5), we conclude that the Kullback-Leibler information of the density indexed by $(\boldsymbol{\theta}^*, \Lambda^*, F^*)$ with respect to the true density is non-positive. Therefore, these two densities must be equal almost surely. For $R = 1$, this implies that

$$P(Y, \Delta | \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}^*, \Lambda^*) F^*(\mathbf{X}, \mathbf{Z}) = P(Y, \Delta | \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) F_0(\mathbf{X}, \mathbf{Z}).$$

It then follows from condition (C.1) that $\boldsymbol{\theta}^* = \boldsymbol{\theta}_0$, $\Lambda^* = \Lambda_0$ and $F^* = F_0$. The consistency is therefore proved.

Proof of Theorem 2

The proof consists of four major steps.

Step 1

Establishing the invertibility of the information operator for $(\boldsymbol{\theta}_0, \Lambda_0, F_0)$ and the existence of the least favorable direction for $\boldsymbol{\theta}_0$.

Let $l_{\boldsymbol{\theta}}$ denote the score for $\boldsymbol{\theta}_0$, $l_{\Lambda}[h_1]$ denote the score along the submodel $\Lambda_0 + \varepsilon \int h_1 d\Lambda_0$, and $l_F[h_2]$ denote the score along the submodel $[1 + \varepsilon\{h_2(\mathbf{x}, \mathbf{z}) - E[h_2(\mathbf{X}, \mathbf{Z})]\}]dF_0(\mathbf{x}, \mathbf{z})$ based on one complete observation $(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$, where $h_1(Y)$ and $h_2(\mathbf{X}, \mathbf{Z})$ belong to the $L_2(P)$ space. To be specific,

$$\begin{aligned} l_{\boldsymbol{\theta}} &= \Delta(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T + (\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \Lambda_0(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0), \\ l_{\Lambda}[h_1] &= \Delta h_1(Y) + \int_0^Y h_1(t) d\Lambda_0(t) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0), \\ l_F[h_2] &= h_2(\mathbf{X}, \mathbf{Z}) - E[h_2(\mathbf{X}, \mathbf{Z})]. \end{aligned}$$

For two-phase studies, the score functions are

$$l_{\theta}^o = Rl_{\theta} + (1-R)E(l_{\theta}|Y, \Delta, \mathbf{Z}, \mathbf{W}), l_{\Lambda}^o[h_1] = Rl_{\Lambda}[h_1] + (1-R)E(l_{\Lambda}[h_1]|Y, \Delta, \mathbf{Z}, \mathbf{W}),$$

$$l_F^o[h_2] = Rl_F[h_2] + (1-R)E(l_F[h_2]|Y, \Delta, \mathbf{Z}, \mathbf{W}).$$

The information operator is $(l_{\theta}^o, l_{\Lambda}^o, l_F^o)^* (l_{\theta}^o, l_{\Lambda}^o, l_F^o)$, where $(l_{\theta}^o, l_{\Lambda}^o, l_F^o)^*$ is the dual operator of $(l_{\theta}^o, l_{\Lambda}^o, l_F^o)$. After tedious calculations, the information operator can be written as the sum of an invertible operator and a compact operator from the space $\mathcal{M} \equiv \mathcal{R}^d \times BV[0, \tau] \times BV(\mathcal{D}_{\mathbf{x}, \mathbf{z}})$ to itself, where d is the dimension of θ , $BV[0, \tau]$ is the space of functions with bounded total variations in $[0, \tau]$, and $BV(\mathcal{D}_{\mathbf{x}, \mathbf{z}})$ is the space of functions with bounded total variation in the support of $(\mathbf{X}, \mathbf{Z})$, denoted as $\mathcal{D}_{\mathbf{x}, \mathbf{z}}$. By Theorem 4.7 of Rudin (1973), the information operator is invertible if it is one to one, or equivalently, the Fisher information along any non-trivial submodel is non-zero.

Suppose that the Fisher information is zero along some submodel

$$[\theta_0 + \varepsilon \mathbf{v}, \Lambda_0 + \varepsilon \int \tilde{h}_1 d\Lambda_0, dF_0(\mathbf{x}, \mathbf{z}) \{1 + \varepsilon \tilde{h}_2(\mathbf{x}, \mathbf{z})\}],$$

where $E\{\tilde{h}_2(\mathbf{X}, \mathbf{Z})\} = 0$. Then the score function along this submodel, i.e.,

$l_{\theta}^{oT} \mathbf{v} + l_{\Lambda}^o[h_1] + l_F^o[h_2]$, is zero. We set $R = 1$ to obtain $l_{\theta}^T \mathbf{v} + l_{\Lambda}[\tilde{h}_1] + l_F[\tilde{h}_2] = 0$ with probability one. That is,

$$\Delta \tilde{h}_1(Y) + \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) \int_0^Y \tilde{h}_1(t) d\Lambda_0(t)$$

$$+ [\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T] \mathbf{v} \{ \Delta + \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) \Lambda_0(Y) \}$$

$$+ \tilde{h}_2(\mathbf{X}, \mathbf{Z}) = 0.$$

We multiply both sides by $\tilde{h}_2(\mathbf{X}, \mathbf{Z})$ and take the expectation. Because $\tilde{h}_2$ is orthogonal to l_{θ} and l_{Λ} , we see that $h_2 = 0$. In the resulting equation, we set $\Delta = 1$ and $Y = 0$ to obtain $[\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T] \mathbf{v} + h_1(0) = 0$. Thus, $\mathbf{v} = \mathbf{0}$. The new equation with $\Delta = 1$ is a homogeneous equation for h_1 with only zero solution, implying that $h_1 = 0$. Therefore, the information operator is invertible. Consequently, there exist h_1 and h_2 such that $(l_{\Lambda}^o, l_F^o)^* (l_{\Lambda}^o, l_F^o)[h_1, h_2] = (l_{\Lambda}^o, l_F^o)^* l_{\theta}$. That is, the least favorable direction for θ_0 exists. It is easy to show that $h_1 \in BV[0, \tau]$ and $h_2(\mathbf{x}, \mathbf{z})$ has bounded q -times derivatives in the support.

Step 2

Deriving the score equation for θ_0 along the least favorable direction.

Let $\mathcal{P}_n$ and $\mathcal{P}$ denote the empirical and true measures, respectively. Since $(\hat{\theta}, \hat{\Lambda}, \hat{F})$ maximizes (3), the derivatives of the log-likelihood function with respect to ε along the submodel $(\hat{\theta} + \varepsilon \mathbf{v}, \hat{\Lambda}, dF)$ for any $\mathbf{v}$, the submodel $(\hat{\theta}, \hat{\Lambda} + \varepsilon \int h_1 d\Lambda, dF)$, and the submodel $(\hat{\theta}, \hat{\Lambda}, dF\{1 + \varepsilon(h_2 - \int h_2 dF)\})$ must be zero. Thus, $(\hat{\theta}, \hat{\Lambda}, \hat{F})$ is the solution to the functional equation $\Phi_n(\hat{\theta}, \hat{\Lambda}, \hat{F}) = 0$, where $\Phi_n(\hat{\theta}, \hat{\Lambda}, \hat{F}) = \Phi_{n1}(\hat{\theta}, \hat{\Lambda}, \hat{F}) - \Phi_{n2}(\hat{\theta}, \hat{\Lambda}, \hat{F}) - \Phi_{n3}(\hat{\theta}, \hat{\Lambda}, \hat{F})$,

$$\begin{aligned}
\Psi_{n1}(\boldsymbol{\theta}, \Lambda, F) &= \mathcal{P}_n \left[R(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \} \right] \\
&+ \mathcal{P}_n \left[(1-R) \int (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \} \times \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) dF(\mathbf{x}, \mathbf{z}) \right], \\
\Psi_{n2}(\boldsymbol{\theta}, \Lambda, F) &= \mathcal{P}_n \left[\Delta h_1(Y) + R \int_0^Y h_1 d\Lambda \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \right] \\
&+ \mathcal{P}_n \left[(1-R) \int_0^Y h_1 d\Lambda \int \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \times \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) dF(\mathbf{x}, \mathbf{z}) \right], \\
\Psi_{n3}(\boldsymbol{\theta}, \Lambda, F) &= \mathcal{P}_n \left[R \frac{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z})/a_n\} \{h_2(\mathbf{X}, \mathbf{Z}_j) - \int h_2(\mathbf{x}, \mathbf{z}) dF(\mathbf{x}, \mathbf{z})\}}{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z})/a_n\}} \right] \\
&+ \mathcal{P}_n \left[(1-R) \int \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) \times \{h_2(\mathbf{x}, \mathbf{z}) - \int h_2(\mathbf{x}, \mathbf{z}) dF(\mathbf{x}, \mathbf{z})\} dF(\mathbf{x}, \mathbf{z}) \right],
\end{aligned}$$

and

$$\mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) = \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) K\{(\mathbf{z} - \mathbf{Z})/a_n\}}{\int \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) K\{(\mathbf{z} - \mathbf{Z})/a_n\} dF(\mathbf{x}, \mathbf{z})}.$$

Let $\tilde{\Phi}_n(\boldsymbol{\theta}, \Lambda, F)$ be the same as $\Phi_n(\boldsymbol{\theta}, \Lambda, F)$ except that $\mathcal{P}_n$ is replaced by $\mathcal{P}$. Clearly, $\hat{\boldsymbol{\theta}}$ satisfies the following equation

$$n^{1/2} \left[\Psi_n(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) - \tilde{\Psi}_n(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) \right] = -n^{1/2} \tilde{\Psi}_n(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}). \quad (\text{A.6})$$

Step 3

Obtaining the asymptotic linear expansion of the score function along the least favorable sub-model.

We wish to show that

$$n^{1/2} \left[\Psi_n(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) - \tilde{\Psi}_n(\hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) \right] = n^{1/2} (\mathcal{P}_n - \mathcal{P}) \left(l_{\boldsymbol{\theta}}^o - l_{\Lambda}^o[h_1] - l_F^o[h_2] \right) + o_p(1). \quad (\text{A.7})$$

We prove this result by using Theorem 2.11.22 of van der Vaart and Wellner (1996). Note that the left-hand side of (A.7) is an empirical process of the following classes of functions indexed by $(\boldsymbol{\theta}, \hat{\Lambda}, \hat{F})$:

$$\begin{aligned}
\mathcal{F}_{1n} &= \left[R(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \} \right. \\
&+ (1-R) \int (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \} \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) dF(\mathbf{x}, \mathbf{z}) : \\
&\quad \left. |\boldsymbol{\theta} - \boldsymbol{\theta}_0| + \|\Lambda - \Lambda_0\| + \|F - F_0\| \leq \varepsilon_0 \right], \\
\mathcal{F}_{2n} &= \left[\Delta h_1(Y) + R \int_0^Y h_1 d\Lambda \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \right. \\
&+ (1-R) \int_0^Y h_1 d\Lambda \int \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}, \Lambda) \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) dF(\mathbf{x}, \mathbf{z}) : \\
&\quad \left. |\boldsymbol{\theta} - \boldsymbol{\theta}_0| + \|\Lambda - \Lambda_0\| + \|F - F_0\| \leq \varepsilon_0 \right],
\end{aligned}$$

and

$$\begin{aligned} \mathcal{F}_{3n} = & \left[R \frac{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z})/a_n\} \{h_2(\mathbf{X}, \mathbf{Z}_j) - \int h_2(\mathbf{x}, \mathbf{z}) dF(\mathbf{x}, \mathbf{z})\}}{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z})/a_n\}} \right] \\ & + (1-R) \int \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}, \Lambda, F) \{h_2(\mathbf{x}, \mathbf{z}) - \int h_2(\mathbf{x}, \mathbf{z}) dF(\mathbf{x}, \mathbf{z})\} dF(\mathbf{x}, \mathbf{z}) : \\ & |\boldsymbol{\theta} - \boldsymbol{\theta}_0| + \|\Lambda - \Lambda_0\| + \|F - F_0\| \leq \varepsilon_0, \end{aligned}$$

where $\|\Lambda - \Lambda_0\|$ is the supreme norm in $[0, \tau]$, and $\|F - F_0\|$ is the supreme norm in $\mathcal{D}_{\mathbf{x}, \mathbf{z}}$. By Theorem 1 and the theory on kernel estimation, it is straightforward to verify that uniformly in $(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W})$,

$$\begin{aligned} & R(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \left\{ \Delta + \hat{\Lambda}(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) \right\} \\ & + (1-R) \int (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \left\{ \Delta + \hat{\Lambda}(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) \right\} \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}, F) dF(\mathbf{x}, \mathbf{z}) \\ & \rightarrow R(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \left\{ \Delta + \Lambda_0(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) \right\} \\ & + (1-R) \int (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \left\{ \Delta + \Lambda_0(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) \right\} \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) dF_0(\mathbf{x}, \mathbf{z})}{\int \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) dF_0(\mathbf{x}, \mathbf{z})} \\ & = Rl_{\boldsymbol{\theta}} + (1-R)E[l_{\boldsymbol{\theta}}|Y, \Delta, \mathbf{Z}, \mathbf{W}] = l_{\boldsymbol{\theta}}^o. \end{aligned}$$

Likewise,

$$\begin{aligned} & \Delta h_1(Y) + R \int_0^Y h_1 d\hat{\Lambda} \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) \\ & + (1-R) \int_0^Y h_1 d\hat{\Lambda} \int \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}) \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}, F) d\hat{F}(\mathbf{x}, \mathbf{z}) \\ & \rightarrow \Delta h_1(Y) + R \int_0^Y h_1 d\Lambda_0 \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) \\ & + (1-R) \int_0^Y h_1 d\Lambda_0 \int \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) dF_0(\mathbf{x}, \mathbf{z})}{\int \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_0, \Lambda_0) dF_0(\mathbf{x}, \mathbf{z})} \\ & = Rl_{\Lambda}[h_1] + (1-R)E[l_{\Lambda}[h_1]|Y, \Delta, \mathbf{Z}, \mathbf{W}] = l_{\Lambda}^o[h_1], \end{aligned}$$

and

$$\begin{aligned} & R \frac{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z})/a_n\} \{h_2(\mathbf{X}, \mathbf{Z}_j) - \int h_2(\mathbf{x}, \mathbf{z}) d\hat{F}(\mathbf{x}, \mathbf{z})\}}{\sum_{j=1}^n K\{(\mathbf{Z}_j - \mathbf{Z})/a_n\}} \\ & + (1-R) \int \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \hat{\boldsymbol{\theta}}, \hat{\Lambda}, \hat{F}) \left\{ h_2(\mathbf{x}, \mathbf{z}) - \int h_2(\mathbf{x}, \mathbf{z}) d\hat{F}(\mathbf{x}, \mathbf{z}) \right\} d\hat{F}(\mathbf{x}, \mathbf{z}) \\ & \rightarrow R[h_2(\mathbf{X}, \mathbf{Z}) - E\{h_2(\mathbf{X}, \mathbf{Z})\}] + (1-R)E[h_2(\mathbf{X}, \mathbf{Z}) - E\{h_2(\mathbf{X}, \mathbf{Z})\}|Y, \Delta, \mathbf{Z}, \mathbf{W}] = l_F^o[h_2]. \end{aligned}$$

Clearly, all the functions in the classes $\mathcal{F}_{kn}$ ($k = 1, 2, 3$) are uniformly bounded. We wish to verify all the conditions in Theorem 2.11.22 of van der Vaart and Wellner (1996). We first show that the classes of functions $\mathcal{F}_{kn}$ ($k = 1, 2, 3$) satisfy the uniform entropy condition in Theorem 2.11.22 of van der Vaart and Wellner (1996). Choose any two functions from $\mathcal{F}_{1n}$, say f_1 and f_2 , which are indexed by $(\boldsymbol{\theta}_1, \Lambda_1, F_1)$ and $(\boldsymbol{\theta}_2, \Lambda_2, F_2)$, respectively. The difference between these two functions is bounded from above by

$$\begin{aligned}
& |\Lambda_1(Y)\mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_1, \Lambda_1) - \Lambda_2(Y)\mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \boldsymbol{\theta}_2, \Lambda_2)| \\
& + \left| \int (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \{ \Delta + \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_1, \Lambda_1) \} \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1, \Lambda_1, F_1) (dF_1 - dF_2)(\mathbf{x}, \mathbf{z}) \right| \\
& + c_1 \int |\mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_1, \Lambda_1) - \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_2, \Lambda_2)| \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1, \Lambda_1, F_1) dF_2(\mathbf{x}, \mathbf{z}) \\
& + c_1 \int (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \boldsymbol{\theta}_2, \Lambda_2) |\mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_1, \Lambda_1, F_1) \\
& \quad - \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \boldsymbol{\theta}_2, \Lambda_2, F_2)| dF_2(\mathbf{x}, \mathbf{z}) \\
& \quad \equiv (i) + (ii) + (iii) + (iv),
\end{aligned}$$

where c_1 is an upper bound for the norm of $(\mathbf{X}, \mathbf{Z}, \mathbf{W})$ in their support. By the mean-value theorem, there exists a constant c_2 such that

$$(i) \leq c_2 \{ |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + |\Lambda_1(Y) - \Lambda_2(Y)| \}.$$

Using the variable transformation, we see that

$$\begin{aligned}
(ii) = & \left| \int_{\mathbf{x}, \tilde{\mathbf{z}}} (\mathbf{x}^T, \mathbf{Z}^T + a_n \tilde{\mathbf{z}}^T, \mathbf{W}^T)^T \{ \Delta + \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}, \mathbf{W}; \boldsymbol{\theta}_1, \Lambda_1) \} \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}, \mathbf{W}; \boldsymbol{\theta}_1, \Lambda_1)}{g_n(\mathbf{Z})} \right. \\
& \left. \times (dF_1 - dF_2)(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) \right|,
\end{aligned}$$

where $g_n(\mathbf{Z}) = \int \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) K(\tilde{\mathbf{z}}) dF_1(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}})$. By applying sequential integration by part and noting that $g_n(\mathbf{Z})$ converges uniformly in $\mathbf{Z}$ to $\int_{\mathbf{x}} \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z}) F_1(d\mathbf{x}, \mathbf{Z})$ and is thus bounded from zero, we obtain

$$(ii) \leq c_3 \int_{\mathbf{x}, \tilde{\mathbf{z}}} |F_1(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) - F_2(\mathbf{x}, \mathbf{Z}) + a_n \tilde{\mathbf{z}}| d\mathbf{x} d\tilde{\mathbf{z}}$$

for some constant c_3 . By the mean-value theorem,

$$(iii) \leq c_4 \{ |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + |\Lambda_1(Y) - \Lambda_2(Y)| \} \int a_n^{-d} K\{(\mathbf{z} - \mathbf{Z})/a_n\} / g_n(\mathbf{Z}) dF_2(\mathbf{x}, \mathbf{z}),$$

which is further bounded by $c_5 \{ |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + |\Lambda_1(Y) - \Lambda_2(Y)| \}$ for some constants c_4 and c_5 . Likewise, there exists some constant c_6 such that

$$(iv) \leq c_6 \left\{ |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + |\Lambda_1(Y) - \Lambda_2(Y)| + \int_{\mathbf{x}, \tilde{\mathbf{z}}} |F_1(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) - F_2(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}})| d\mathbf{x} d\tilde{\mathbf{z}} \right\}.$$

Therefore,

$$|f_1 - f_2| \leq c_7 \left\{ |\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2| + |\Lambda_1(Y) - \Lambda_2(Y)| + \int_{\mathbf{x}, \tilde{\mathbf{z}}} |F_1(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) - F_2(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}})| d\mathbf{x} d\tilde{\mathbf{z}} \right\}$$

for some constant c_7 . In addition, the integration in the above inequality is over the domain of the support of $\mathbf{X}$ and the support of the kernel function. Thus, the Cauchy-Schwarz inequality implies that, for any finite measure Q ,

$$\|f_1 - f_2\|_{L_2(Q)} \leq c_7 \left[|\theta_1 - \theta_2| + \|\Lambda_1(Y) - \Lambda_2(Y)\|_{L_2(Q)} + \left\{ \int_{\mathbf{x}, \tilde{\mathbf{z}}} \|F_1(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) - F_2(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}})\|_{L_2(Q)}^2 d\mathbf{x} d\tilde{\mathbf{z}} \right\}^{1/2} \right].$$

Since $\tilde{\mathbf{z}}$ is bounded and a_n is very small, we can expand the measure Q to the space

$$\{\mathbf{Z} + a_n \tilde{\mathbf{z}} : \mathbf{Z} \in \mathcal{D}_{\mathbf{Z}}, \tilde{\mathbf{z}} \text{ in the support of } K(\cdot)\}$$

and still obtain a finite measure Q . Thus,

$$\int_{\mathbf{x}, \tilde{\mathbf{z}}} \|F_1(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}}) - F_2(\mathbf{x}, \mathbf{Z} + a_n \tilde{\mathbf{z}})\|_{L_2(Q)}^2 d\mathbf{x} d\tilde{\mathbf{z}} \leq \int_{\mathbf{x}, \tilde{\mathbf{z}}} \|F_1(\mathbf{x}, \mathbf{Z}) - F_2(\mathbf{x}, \mathbf{Z})\|_{L_2(\tilde{Q})}^2 d\mathbf{x} d\tilde{\mathbf{z}}.$$

It follows that

$$\|f_1 - f_2\|_{L_2(Q)} \leq c_7 \left\{ |\theta_1 - \theta_2| + \|\Lambda_1(Y) - \Lambda_2(Y)\|_{L_2(Q)} + \|F_1(\mathbf{X}, \mathbf{Z}) - F_2(\mathbf{X}, \mathbf{Z})\|_{L_2(\tilde{P} \times \tilde{Q})} \right\}, \quad (\text{A.8})$$

where $\tilde{P}$ is the uniform measure in $\mathcal{D}_{\mathbf{x}}$. We conclude that

$$N((\varepsilon, \mathcal{F}_{1n}, L_2(Q))) \leq c_8 N\{\varepsilon/3, (\theta: |\theta - \theta_0| < \varepsilon_0), |\cdot|\} N\{\varepsilon/3, (\Lambda: \|\Lambda - \Lambda_0\|_{\infty} < \varepsilon_0), L_2(Q)\} \times N\{\varepsilon/3, (F: \|F - F_0\|_{\infty} < \varepsilon_0), L_2(\tilde{P} \times \tilde{Q})\}, \quad (\text{A.9})$$

where c_8 is a constant, and $N(\cdot, \cdot, \cdot)$ denotes the covering function. On the right-hand side of (A.9), the first covering number is $O(1/\varepsilon^d)$, the second covering number is $O\{\exp(1/\varepsilon)\}$ according to Theorem 2.7.5 of van der Vaart and Wellner (1996), and the third covering number is $O\{\exp(1/\varepsilon^V)\}$ for some positive index V . The last result is due to the fact that $(F: \|F - F_0\|_{\infty} < \varepsilon)$ is in the symmetric convex hull of a VC-class $[I\{\mathbf{a} < (\mathbf{X}^T, \mathbf{Z}^T)^T \mathbf{b}\} : \mathbf{a}, \mathbf{b} \in \mathcal{R}^{d_{\mathbf{x}} + d_{\mathbf{z}}}]$ and follows from Theorem 2.6.9 of van der Vaart and Wellner (1996). Therefore, (A.9) implies that $\mathcal{F}_{1n}$ satisfies the uniform entropy condition in Theorem 2.11.22 of van der Vaart and Wellner (1996). By the same arguments, $\mathcal{F}_{2n}$ and $\mathcal{F}_{3n}$ also satisfy the uniform entropy condition.

If we replace measure Q by measure P , then (A.8) implies that the functions in the classes $\mathcal{F}_{kn}$ ($k = 1, 2, 3$) are Lipschitz continuous with respect to (θ, Λ, F) in the metric defined as

$$\rho((\theta_1, \Lambda_1, F_1), (\theta_2, \Lambda_2, F_2)) = |\theta_1 - \theta_2| + \|\Lambda_1 - \Lambda_2\|_{L_2(P)} + \|F_1 - F_2\|_{L_2(P)}.$$

As a result, condition (2.11.21) in Theorem 2.11.22 of van der Vaart and Wellner (1996) holds. In addition, the total boundedness of the index set (θ, Λ, F) holds due to the pre-compactness of (θ, Λ, F) under the uniform metric.

We have now verified all the conditions in Theorem 2.11.22 of van der Vaart and Wellner (1996). Thus, equation (A.7) follows from that theorem.

Step 4

Proving the asymptotic normality and efficiency of $\hat{\theta}$.

In light of equations (A.6) and (A.7),

$$-n^{1/2} \left\{ \tilde{\Psi}_{n1}(\hat{\theta}, \hat{\Lambda}, \hat{F}) - \tilde{\Psi}_{n2}(\hat{\theta}, \hat{\Lambda}, \hat{F}) - \tilde{\Psi}_{n3}(\hat{\theta}, \hat{\Lambda}, \hat{F}) \right\} = n^{1/2} (\mathcal{D}_n - \mathcal{D}) \left(l_{\theta}^o - l_{\Lambda}^o[h_1] - l_F^o[h_2] \right) + o_p(1). \quad (\text{A.10})$$

The left-hand side of (A.10) can be linearized around $(\theta_0, \Lambda_0, F_0)$. Specifically,

$$\begin{aligned} \tilde{\Psi}_{n1}(\hat{\theta}, \hat{\Lambda}, \hat{F}) &= \tilde{\Psi}_{n1}(\theta_0, \Lambda_0, F_0) + \mathcal{D} \left[R(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \Lambda(Y) \frac{\partial}{\partial \theta} \{ \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta^*, \Lambda^*) \} (\hat{\theta} - \right. \\ &+ \frac{\partial}{\partial \Lambda} \{ \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta^*, \Lambda^*) \} \{ \hat{\Lambda}(Y) - \Lambda_0(Y) \} \Big] \\ &+ \mathcal{D} \left[(1-R) f(\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \frac{\partial}{\partial \theta} \{ \Delta + \Lambda^*(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta^*, \Lambda^*) \times \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \theta^*, \Lambda^*, F^*) \} (\hat{\theta} - \theta_0) d\hat{F}(\mathbf{x}, \mathbf{z}) \right] \\ &+ \mathcal{D} \left[(1-R) f(\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \frac{\partial}{\partial \Lambda} \{ \Delta + \Lambda^*(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta^*, \Lambda^*) \times \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \theta^*, \Lambda^*, F^*) \} (\hat{\Lambda}(Y) - \Lambda_0(Y)) d\hat{F}(\mathbf{x}, \right. \\ &+ \mathcal{D} \left[(1-R) f(\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \frac{\partial}{\partial F} \{ \Delta + \Lambda^*(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta^*, \Lambda^*) + \mathcal{B}_{4n}(Y, \Delta, \mathbf{W}, \mathbf{Z}, \mathbf{x}, \mathbf{z}; \theta^*, \Lambda^*, F^*) dF^*(\mathbf{x}, \mathbf{z}) \} [\hat{F} - F_0] \right]. \end{aligned}$$

where $\frac{\partial}{\partial \theta}$, $\frac{\partial}{\partial \Lambda}$ and $\frac{\partial}{\partial F}$ denote pathwise derivatives, and $(\theta^*, \Lambda^*, F^*)$ lies between $(\hat{\theta}, \hat{\Lambda}, \hat{F})$ and $(\theta_0, \Lambda_0, F_0)$. Similar expansions can be obtained for Φ_{n2} and Φ_{n3} . By the kernel approximation, we can show that the left-hand side of (A.10) is equal to

$$\begin{aligned} &n^{1/2} (1 + o_p(1)) \left\{ E \left(l_{\theta\theta}^o (\hat{\theta} - \theta_0) + l_{\theta\Lambda}^o [\hat{\Lambda} - \Lambda_0] + l_{\theta F}^o [\hat{F} - F_0] \right) \right. \\ &- E \left(l_{\Lambda\theta}^o [h_1] (\hat{\theta} - \theta_0) + l_{\Lambda\Lambda}^o [h_1, \hat{\Lambda} - \Lambda_0] + l_{\Lambda F}^o [h_1, \hat{F} - F_0] \right) \\ &- E \left(l_{F\theta}^o [h_2] (\hat{\theta} - \theta_0) + l_{F\Lambda}^o [h_2, \hat{\Lambda} - \Lambda_0] + l_{FF}^o [h_2, \hat{F} - F_0] \right) \Big\} \\ &+ n^{1/2} \left\{ \tilde{\Psi}_{n1}(\theta_0, \Lambda_0, F_0) - \tilde{\Psi}_{n2}(\theta_0, \Lambda_0, F_0) - \tilde{\Psi}_{n3}(\theta_0, \Lambda_0, F_0) \right\}, \quad (\text{A.11}) \end{aligned}$$

where $l_{\theta\theta}^o$ is the second derivative of l^o with respect to θ , $l_{\theta\Lambda}^o[h]$ and $l_{\theta F}^o[h]$ are the derivatives of l_{θ}^o with respect to Λ and F along the direction h , $l_{\Lambda\Lambda}^o[h_1, h_2]$ and $l_{\Lambda F}^o[h_1, h_2]$ are the derivatives of $l_{\Lambda}^o[h_1]$ with respect to Λ and F along the direction h_2 , and $l_{FF}^o[h_1, h_2]$ are the derivatives of $l_F^o[h_1]$ with respect to Λ and F along the direction h_2 .

Since h_1 and h_2 are chosen to be the least favorable directions for θ_0 given at the end of Step 1, we have $E(l_{\Lambda\Lambda}^o[h_1, \hat{\Lambda} - \Lambda_0] + l_{\Lambda F}^o[h_1, \hat{F} - F_0]) = E(l_{\theta\Lambda}^o[\hat{\Lambda} - \Lambda_0])$ and

$E(l_{F\Lambda}^o[h_2, \hat{\Lambda} - \Lambda_0] + l_{FF}^o[h_2, \hat{F} - F_0]) = E(l_{\theta F}^o[\hat{F} - F_0])$. Thus, the first term in (A.11) is

equivalent to $n^{1/2} \Sigma(\hat{\theta} - \theta_0)$, where $\Sigma = -E(l_{\theta\theta}^o) + E(l_{\Lambda\theta}^o[h_1]) + E(l_{F\theta}^o[h_2])$, which is an invertible matrix due to the invertibility of the information operator for $(\theta_0, \Lambda_0, F_0)$. Finally, since

$$\mathcal{D} [R(\mathbf{X}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{X}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) \}] = 0$$

and

$$\mathcal{P} \left[(1-R) \int (\mathbf{x}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) \} \times \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) F_0(d\mathbf{x}, \mathbf{Z})}{\int_{\mathbf{x}} \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) F_0(d\mathbf{x}, \mathbf{Z})} \right] = \mathbf{0},$$

we have

$$\begin{aligned} \tilde{\Psi}_{n1}(\theta_0, \Lambda_0, F_0) = & \mathcal{P} \left[(1-R) \int_{\mathbf{x}, \mathbf{z}} (\mathbf{x}^T, \mathbf{z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta_0, \Lambda_0) \} \right. \\ & \times \left. \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta_0, \Lambda_0) K\{(\mathbf{z}-\mathbf{Z})/a_n\} dF_0(\mathbf{x}, \mathbf{z})}{\int_{\mathbf{x}} \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{z}, \mathbf{W}; \theta_0, \Lambda_0) K\{(\mathbf{z}-\mathbf{Z})/a_n\} dF_0(\mathbf{x}, \mathbf{z})} \right] \\ & - \mathcal{P} \left[(1-R) \int (\mathbf{x}^T, \mathbf{Z}^T, \mathbf{W}^T)^T \{ \Delta + \Lambda(Y) \mathcal{B}_1(Y, \Delta, \mathbf{x}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) \} \right. \\ & \times \left. \frac{\mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) F_0(d\mathbf{x}, \mathbf{Z})}{\int_{\mathbf{x}} \mathcal{B}_2(Y, \Delta, \mathbf{x}, \mathbf{Z}, \mathbf{W}; \theta_0, \Lambda_0) F_0(d\mathbf{x}, \mathbf{Z})} \right]. \end{aligned} \quad (\text{A.12})$$

Condition (C.4) implies that, for any q -times differentiable function $g(\mathbf{z})$,

$$a_n^{-d_{\mathbf{z}}} \int g(\mathbf{z}) K\{(\mathbf{z}-\tilde{\mathbf{z}})/a_n\} d\mathbf{z} - g(\tilde{\mathbf{z}}) = \int \{g(\tilde{\mathbf{z}} + a_n \mathbf{z}) - g(\tilde{\mathbf{z}})\} K(\mathbf{z}) d\mathbf{z} = O(a_n^d).$$

Thus, (A.12) is $O(n^{1/2} a_n^q)$, which is $o_p(1)$ by condition (C.5). The same arguments apply to $\Phi_{n2}(\theta_0, \Lambda_0, F_0)$ and $\Phi_{n3}(\theta_0, \Lambda_0, F_0)$. Thus, the last term of (A.11) is $o_p(1)$.

It follows from (A.10) that

$$n^{1/2}(1+o_p(1)) \sum (\hat{\theta} - \theta_0) + o_p(1) = n^{1/2}(\mathcal{P}_n - \mathcal{P}) \left(l_{\theta}^o - l_{\Lambda}^o[h_1] - l_F^o[h_2] \right).$$

This proves the asymptotic normality in Theorem 2. Since $\sum^{-1} (l_{\theta}^o - l_{\Lambda}^o[h_1] - l_F^o[h_2])$ is the efficient influence function for θ_0 , its limiting covariance matrix attains the semiparametric efficiency bound.

Given the existence of the least favorable directions, we can construct the least favorable submodel. We proceed to verify the conditions of Murphy and van der Vaart (2000) and conclude that the inverse of the negative Hessian matrix of the profile likelihood function is a consistent estimator for the limiting covariance matrix of $n^{1/2}(\hat{\theta} - \theta_0)$.

Table 1

Summary Statistics for Simulation Studies on the Proposed Methods

	Covariate	Proportional hazards model				Proportional odds model			
		Bias	SE	SEE	CP	Bias	SE	SEE	CP
Discrete Z	X	-0.003	0.072	0.072	0.949	-0.004	0.085	0.083	0.944
	Z	0.003	0.112	0.116	0.953	0.004	0.138	0.135	0.944
	W	0.001	0.071	0.072	0.955	0.001	0.088	0.092	0.960
Continuous Z	X	0.005	0.276	0.273	0.946	0.006	0.319	0.309	0.941
	Z	-0.014	0.358	0.355	0.947	-0.015	0.422	0.412	0.944
	W	0.005	0.169	0.168	0.948	0.007	0.204	0.210	0.952

NOTE: Bias and SE are the bias and standard error of the parameter estimator, SEE is the mean of the standard error estimator, and CP is the coverage probability of the 95% confidence interval. Each entry is based on 10,000 replicates.

Table 2

Summary Statistics for Simulation Studies on the Existing Methods

Covariate	Full Cohort		Borgan et al.		Zeng et al.	
	Bias	SE	Bias	SE	Bias	SE
X	0.001	0.225	0.001	0.289	0.189	0.205
Z	-0.007	0.327	0.005	0.422	-0.368	0.244
W	0.047	0.169	0.010	0.217	0.055	0.169

NOTE: Bias and SE are the bias and standard error of the parameter estimator. Each entry is based on 10,000 replicates.

Table 3

Analysis of Wilms' Tumor Data

Parameter	Proportional hazards model				Proportional odds model			
	Full-cohort		Case-cohort		Full-cohort		Case-cohort	
	Est	SE	Est	SE	Est	SE	Est	SE
UH	4.041	0.413	4.401	0.439	5.911	0.549	6.204	0.637
UH*Age1	-2.634	0.464	-2.890	0.505	-4.323	0.618	-4.585	0.727
UH*Age2	-0.058	0.034	-0.065	0.040	-0.046	0.043	0.050	0.052
Age1	-0.661	0.326	-0.593	0.335	-0.721	0.246	-0.665	0.318
Age 2	0.104	0.017	0.107	0.016	0.111	0.018	0.113	0.018
Stage	1.346	0.244	1.149	0.247	1.556	0.268	1.421	0.287
Diameter	0.069	0.014	0.064	0.015	0.075	0.014	0.072	0.016
Stage*Diameter	-0.076	0.019	-0.066	0.019	-0.088	0.021	-0.083	0.022

NOTE: UH denotes unfavorable central histology. Age1 and Age2 indicate age < 1 year and age 1 year. Est and SE are the mean estimated parameter and mean estimated standard error over 1000 simulated subcohorts.