The Canadian Journal of Statistics
Vol. 39, No. 4, 2011, Pages 557–577
La revue canadienne de statistique

Relative risk regression for current status data in case-cohort studies

Zhiguo LI[1] and Bin NAN[2]*

[1] *Department of Biostatistics & Bioinformatics, Duke University, Durham, NC, USA*
[2] *Department of Biostatistics, University of Michigan, Ann Arbor, MI, USA*

Key words and phrases: Estimated Weights; interval censoring; missing covariates; two-phase design; weighted bootstrap; weighted likelihood

MSC 2010: Primary 62E20, 62N02; secondary 62D05.

Abstract: We propose using the weighted likelihood method to fit a general relative risk regression model for the current status data with missing data as arise, for example, in case-cohort studies. The missingness probability is either known or can be reasonably estimated. Asymptotic properties of the weighted likelihood estimators are established. For the case of using estimated weights, we construct a general theorem that guarantees the asymptotic normality of the M-estimator of a finite dimensional parameter in a class of semiparametric models, where the infinite dimensional parameter is allowed to converge at a slower than parametric rate, and some other parameters in the objective function are estimated a priori. The weighted bootstrap method is employed to estimate the variances. Simulations show that the proposed method works well for finite sample sizes. A motivating example of the case-cohort study from an HIV vaccine trial is used to demonstrate the proposed method. *The Canadian Journal of Statistics* 39: 557–577; 2011 © 2011 Statistical Society of Canada

Résumé: Nous proposons d'utiliser la méthode de vraisemblance pondérée pour ajuster un modèle de régression général pour le risque relatif sur des données de statut présent avec données man-quantes. Une telle situation se produit dans les études cas-cohorte. La probabilité d'être manquante est connue ou bien elle peut être estimée de façon raisonnable. Les propriétés asymptotiques des estimateurs de vraisemblance pondérée sont obtenues. Lorsque des poids estimés sont utilisés, nous obtenons un théorème général garantissant la normalité asymptotique du M-estimateur d'un pa-ramètre de dimension fini appartenant à une classe de modèles semi-paramétriques, pour laquelle le paramètre de dimension infinie peut converger à un taux plus lent que le taux paramétrique, et que d'autres paramètres de la fonction objective sont estimés *a priori* La méthode d'auto-amorçage pondérée est utilisée pour estimer les variances. Des simulations montrent que la méthode proposée fonctionne bien pour de petits échantillons. Une étude cas-cohorte provenant d'un essai clinique sur un vaccin contre le VIH/sida sert à motiver la méthodologie proposée. *La revue canadienne de statistique* 39: 557–577; 2011 © 2011 Société statistique du Canada

1. INTRODUCTION

The case-cohort design, originally proposed by Prentice (1986), is a cost-effective approach in conducting large epidemiologic studies in which the outcome of interest is time to event and some covariates are difficult or expensive to measure. In such a study design, these covariates are only measured for all the subjects who have experienced the event of interest and a random subsample

* *Author to whom correspondence may be addressed.*
E-mail: bnan@umich.edu

of the entire cohort. Statistical inference with data from case-cohort studies must take the missing covariates into account.

There is a rich literature of statistical methodology in analyzing the case-cohort data. Among many others, Prentice (1986) and Self & Prentice (1988) studied the relative risk model that includes the Cox model (Cox, 1972) as a special example, Kulich & Lin (2000) studied the additive hazards model, Lu & Tsiatis (2006) and Kong, Cai, & Sen (2006) studied the transformation model, Nan, Yu, & Kalbfleisch (2006) and Nan, Kalbfleisch, & Yu (2009) studied the accelerated failure time model, and Nan (2004) and Nan, Emond, & Wellner (2004) studied the semiparametric efficient estimation for case-cohort studies. All of these methods primarily focus on right censored data. Often in practice, particularly in HIV studies, however, the event time is interval censored, that is, the event time for a subject falls into some random time interval. Gilbert et al. (2005) analyzed interval censored case-cohort data by approximating the event time as to be right censored. Clearly such approximation can cause biased parameter estimation. The only work we are aware of, which directly attacks the interval censoring mechanism in case-cohort studies, is by Li, Gilbert, & Nan (2008) who considered the Cox model and particularly assumed that the inspection time intervals are *fixed*, thus the model is *parametric*.

In this article, we consider a family of *semiparametric* regression models for the current status data in two-phase sampling designs (Neyman, 1938) that include case-cohort studies as special examples. Current status data are a special type of interval censored data in which the inspection time intervals are *random* in contrast to fixed inspection time intervals, for the latter a parametric model can be fitted. The current status data are also called the "case 1" interval censored data in the literature, in which we only know whether the failure event has occurred or not prior to a random inspection time. The fact that the exact time to event is never observed leads to a $n^{1/3}$ convergence rate for the maximum likelihood estimator of the marginal event time distribution (Groeneboom & Wellner, 1992) and for the baseline cumulative hazard function estimator in the Cox model (Huang, 1996; Murphy & van der Vaart, 2000; van der Vaart, 2002) when there is no missing data. The log hazard ratio estimator in the Cox model, however, still converges with $\sqrt{n}$ rate and is asymptotically normal and semiparametrically efficient. The model we consider in this article is a general relative risk regression model studied by Prentice & Self (1983) and Thomas (1981) who argued, among others, that in many epidemiologic studies the relative risk is not exponential as what the Cox model assumes, and it is more appropriate to consider other types of relative risk models, for example, a linear relative risk form. We are not aware of any existing work for the relative risk regression with current status data, particularly when covariates are not always observed.

Statistical inference for current status data with missing covariates using the usual nonparametric likelihood approach can be very difficult if not impossible. The weighted likelihood method, however, can be easily applied. One can either maximize the inverse probability weighted log likelihood function (e.g., Kalbfleisch & Lawless, 1988; Skinner et al., 1989), or equivalently solve the weighted score equation (e.g., Manski & Lerman, 1977) to estimate the unknown parameters. When the weighted likelihood approach is applied to parametric models, the asymptotic properties of the regular estimators with $\sqrt{n}$ convergence rate follow readily from the results for M-estimation (e.g., van der Vaart, 1998). In a recent work on semiparametric models for two-phase sampling designs in which the infinite dimensional nuisance parameter can be estimated at $\sqrt{n}$ rate, Breslow & Wellner (2007) considered the weighted likelihood method and derived asymptotic results for both Bernoulli sampling and finite population stratified sampling in selecting the phase two sample. Their approach, however, does not apply when the convergence rate of the nuisance parameter estimator is slower than $\sqrt{n}$, which is indeed the case for the current status data with missing covariates as we show later in this article. To solve this problem, particularly when estimated weights are involved in the weighted likelihood, we construct a general theorem that generalizes Theorem 6.1 in Wellner & Zhang (2007), which was developed for their pseudo likelihood method, and then apply the theorem to show that our proposed estimators of the relative risk parameter are asymptotically normal and that using estimated weights improves efficiency. We also provide a different proof of consistency to Huang (1996) where his application of Hoeffding's inequality is incorrect.

The construction of the paper is as follows. In The Weighted Likelihood Estimator Section we provide an algorithm that is a modification of the one given in Huang (1996) for computing the weighted likelihood estimates. In Asymptotic Properties Section we establish the asymptotic properties of the weighted likelihood estimates. We discuss the variance estimation using weighted bootstrap in Variance Estimation Section, and conduct simulations and analyze the data from a case-cohort HIV vaccine study in Numerical Results Section. A brief discussion is given in Discussion Section. In Appendix A, we introduce a general theorem for the proof of asymptotic normality for the weighted likelihood estimates using estimated weights. The proofs of asymptotic properties are provided in Appendix B.

2. THE WEIGHTED LIKELIHOOD ESTIMATOR

Suppose the failure time T follows a relative risk regression model:

$$\Lambda(t|Z) = \Lambda(t)r(\beta^T Z),$$

where $\Lambda(t|Z)$ is the conditional cumulative hazard function of T given Z, $\Lambda(t)$ is the baseline cumulative hazard function, and $r(\cdot)$ is a fixed positive and twice continuously differentiable function. A particularly interesting functional form for $r(\cdot)$ is the linear function: $r(x) = 1 + x$ (Prentice & Self, 1983), as an alternative to the exponential function $r(x) = e^x$ that yields the proportional hazards model originally proposed by Cox (1972) for right-censored data.

In current status data, T is never observed. Instead, an inspection time Y is observed, which is assumed to be independent of T given covariate Z, and it is also known whether the event has happened before Y. We consider the case where the covariate Z can be missing as arise, for example, in case-cohort studies. Let $\Delta = I(T \leq Y)$ where $I(\cdot)$ is the indicator function. Denote the probability of observing Z by $\pi_\alpha(\Delta, V)$, which may depend on a parameter α, the failure status Δ, and an auxiliary variable V that is observed for everyone. For example, in a case-cohort design with stratified sampling for the subcohort, the probability of observing covariate Z is $\pi_\alpha(\Delta, V) = \Delta + (1-\Delta)\sum_{j=1}^{J} p_j I(V \in \mathcal{V}_j)$, where $\mathcal{V}_1, \dots, \mathcal{V}_J$ are J strata determined by the value of the auxiliary variable V, $\alpha = (p_1, \dots, p_J)^T$, and p_j is the probability that a subject is sampled into the subcohort from stratum j, $1 \leq j \leq J$. The parameter α may or may not be known. Later we shall discuss the effect of estimating α from observed data. It is possible that V is part of Z. The density of a single observation $X \equiv (\Delta, Y, Z, V)$ at $x \equiv (\delta, y, z, v)$ can be written as

$$p_{\beta,\Lambda}(x) = \left\{1-\exp\left(-\Lambda(y)r(\beta^T z)\right)\right\}^{\delta}\left\{\exp\left(-\Lambda(y)r(\beta^T z)\right)\right\}^{1-\delta} f(y, z, v), \tag{1}$$

where $f(y, z, v)$ is the joint density of (Y, Z, V). The parameter of interest is β, and $\Lambda(\cdot)$ is a nuisance parameter.

Let $X_1, \dots, X_n$ be n independent and identically distributed (i.i.d.) copies of X. The complete data log likelihood function, up to an additive constant, is

$$\begin{aligned} l_n(\beta, \Lambda) &= \sum_{i=1}^{n} l(\beta, \Lambda; X_i) \\ &= \sum_{i=1}^{n} \left[\Delta_i \log\left\{1-\exp\left(-\Lambda(Y_i)r(\beta^T Z_i)\right)\right\}-(1-\Delta_i)\Lambda(Y_i)r(\beta^T Z_i)\right]. \end{aligned} \tag{2}$$

This is the likelihood function studied in Huang (1996) when $r(\cdot) = \exp(\cdot)$.

Because Z_i's are only observed for a subsample, the nonparametric maximum likelihood method can be too complicated to be useful. However, we can use the following weighted version of the log likelihood function

$$l_n^w(\beta, \Lambda) = \sum_{i=1}^{n} w_i \left[\Delta_i \log\left\{1 - \exp\left(-\Lambda(Y_i) r(\beta^T Z_i)\right)\right\} - (1-\Delta_i)\Lambda(Y_i) r(\beta^T Z_i)\right], \qquad (3)$$

where $w_i = \xi_i / \pi_\alpha(\Delta_i, V_i)$ with $\xi_i = 1$ if Z_i is observed and 0 otherwise, $1 \leq i \leq n$. For simplicity, here and in the sequel we suppress the dependence of w on α, Δ, and V, except in Estimation with Estimated Weights Section and Variance Estimation Section, where we discuss the weighted likelihood estimator with estimated weights. Note that when α takes its true value, weights w_i have unit expectations, but they do not necessarily sum to n no matter α is estimated or not. The weighted likelihood estimator of the true parameter (β_0, Λ_0) is defined as the maximizer of the weighted log likelihood function (3) with discretized Λ at observed time points and denoted by $(\hat{\beta}_n, \hat{\Lambda}_n)$, that is,

$$\left(\hat{\beta}_n, \hat{\Lambda}_n\right) = \operatorname{argmax} \sum_{i=1}^{n} w_i l(\beta, \Lambda; X_i).$$

Due to the similarity between (2) and (3), a similar algorithm as in Huang (1996) can be developed to obtain $(\hat{\beta}_n, \hat{\Lambda}_n)$ with a general relative risk function r. Let $\left(Y_{(1)}, \ldots, Y_{(n)}\right)$ be the order statistics of $(Y_1, \ldots, Y_n)$. Let $\Delta_{(i)}, Z_{(i)}$, and $w_{(i)}$ be the values of Δ, Z, and w associated with $Y_{(i)}$, $1 \leq i \leq n$. Consider the estimator $\hat{\Lambda}_n(\cdot)$ to be a right-continuous step function on $[0, Y_{(n)}]$ with jumps at $Y_{(i)}$'s and $\hat{\Lambda}_n(0) = 0$. To ensure a bounded and unique estimator $\hat{\Lambda}_n(\cdot)$, we assume that

$$\Delta_{(1)} = 1, \quad \Delta_{(n)} = 0. \qquad (4)$$

Replacing Λ by its estimator $\hat{\Lambda}_n$, we obtain the following score equation for β by differentiating the objective function (3) with respect to β and setting the derivative to 0:

$$\sum_{i=1}^{n} w_{(i)} \left\{ \Delta_{(i)} \frac{\exp\left(-\hat{\Lambda}_n\left(Y_{(i)}\right) r\left(\hat{\beta}_n^T Z_{(i)}\right)\right)}{1 - \exp\left(-\hat{\Lambda}_n\left(Y_{(i)}\right) r\left(\hat{\beta}_n^T Z_{(i)}\right)\right)} - \left(1 - \Delta_{(i)}\right) \right\} \hat{\Lambda}_n\left(Y_{(i)}\right) \dot{r}\left(\hat{\beta}_n^T Z_{(i)}\right) Z_i = 0, \qquad (5)$$

where $\dot{r}(\cdot)$ denotes the derivative of $r(\cdot)$.

Due to the monotonicity constraint on $\hat{\Lambda}_n$, there is no such a simple score equation for $\hat{\Lambda}_n$. However, analogous to Groeneboom & Wellner (1992), $\hat{\Lambda}_n$ can be characterized by a set of inequalities at k_n distinct inspection times $Y_1^* < Y_2^* < \cdots < Y_{k_n}^*$ and an equality as follows:

$$\sum_{Y_j \geq Y_i^*} w_j r\left(\hat{\beta}_n^T Z_j\right) \left\{ \Delta_j \frac{\exp\left(-\hat{\Lambda}_n\left(Y_j\right) r\left(\hat{\beta}_n^T Z_j\right)\right)}{1 - \exp\left(-\hat{\Lambda}_n\left(Y_j\right) r\left(\hat{\beta}_n^T Z_j\right)\right)} - \left(1 - \Delta_j\right) \right\} \leq 0, \qquad (6)$$

for $i = 1, 2, \ldots, k_n$, and

$$\sum_{i=1}^{n} w_i r\left(\hat{\beta}_n^T Z_i\right) \hat{\Lambda}_n(Y_i) \left\{ \Delta_i \frac{\exp\left(-\hat{\Lambda}_n(Y_i) r\left(\hat{\beta}_n^T Z_i\right)\right)}{1 - \exp\left(-\hat{\Lambda}_n(Y_i) r\left(\hat{\beta}_n^T Z_i\right)\right)} - (1 - \Delta_i) \right\} = 0. \qquad (7)$$

This result is an extension of Theorem 2.1 of Huang (1996) and can be derived in a similar way as that of Proposition 1.1 of Groeneboom & Wellner (1992). Detailed calculation is thus omitted here.

Equations (6) and (7) lead to an iterative algorithm to compute $\hat{\Lambda}_n(\cdot, \beta)$ for any fixed β. This algorithm is more efficient than the pool adjacent violators algorithm (Robertson, Wright, & Dykstra, 1988). Define

$$W_\Lambda\left(Y_i^*\right) = \sum_{Y_j \le Y_i^*} w_j r\left(\beta^T Z_j\right)\left\{\Delta_j \frac{\exp\left(-\Lambda(Y_j) r\left(\beta^T Z_j\right)\right)}{1-\exp\left(-\Lambda(Y_j) r\left(\beta^T Z_j\right)\right)} - (1-\Delta_j)\right\}, \tag{8}$$

$$G_\Lambda\left(Y_i^*\right) = \sum_{j=1}^{i} \Delta G_\Lambda\left(Y_j^*\right), \tag{9}$$

with

$$\Delta G_\Lambda\left(Y_j^*\right) = \sum_{Y_k = Y_j^*} w_k r(\beta^T Z_k)\left\{\Delta_k \frac{r\left(\beta^T Z_k\right)\exp\left(-\Lambda(Y_k) r\left(\beta^T Z_k\right)\right)}{\left(1-\exp\left(-\Lambda(Y_k) r\left(\beta^T Z_k\right)\right)^2\right)} + \frac{1-\Delta_k}{\Lambda(Y_k)}\right\} \tag{10}$$

and

$$V_\Lambda\left(Y_i^*\right) = W_\Lambda\left(Y_i^*\right) + \sum_{Y_j^* \le Y_i^*} \Lambda\left(Y_j^*\right) \Delta G_\Lambda\left(Y_j^*\right). \tag{11}$$

Here we add the quantity $w_k r(\beta^T Z_k)(1-\Delta_k)/\Lambda(Y_k)$ to the original definition of $\Delta G_\Lambda(\cdot)$ in Huang (1996, p. 545) to make $\Delta G_\Lambda\left(Y_j^*\right) \equiv G_\Lambda\left(Y_j^*\right) - G_\Lambda\left(Y_{j-1}^*\right) > 0$ with $G_\Lambda(Y_0^*) \equiv 0$, $1 \le j \le n$, a required condition for the algorithm. In fact, the function $G_\Lambda(\cdot)$ above can be chosen arbitrarily as long as $\Delta G_\Lambda(Y_i^*) > 0$, $1 \le i \le n$, and the constructed $V_\Lambda(\cdot)$ is non-decreasing. The point is clearly seen in the proof of Proposition 1.4 and Remark 1.4 of Groeneboom & Wellner (1992). The choices in Groeneboom & Wellner (1992) are based on a second-order Taylor expansion of the log likelihood function, which work well for the nonparametric estimation of the marginal distribution function of T, but numerical issue arises in the semiparametric regression case since their choices of $G_\Lambda(\cdot)$ does not include the second term in the brackets in (10) and thus has zero increments at all inspection times for censored subjects. This problem is resolved by adding a positive quantity to the increments of $G_\Lambda(\cdot)$ at those time points as in (10). Such added quantity also makes $V_\Lambda(\cdot)$ non-decreasing.

Following the proof of Proposition 1.4 of Groeneboom & Wellner (1992), for any fixed β, by using (6) and (7) it can be shown that $\hat{\Lambda}_n(\cdot;\beta)$ maximizes $l_n^w(\beta, \Lambda)$ if and only if $\hat{\Lambda}_n(\cdot;\beta)$ is the left derivative of the greatest convex minorant of the cumulative sum diagram defined by the points (0, 0) and

$$\left(G_{\hat{\Lambda}_n(\cdot,\beta)}\left(Y_i^*\right), V_{\hat{\Lambda}_n(\cdot,\beta)}\left(Y_i^*\right)\right), \quad 1 \le i \le k_n. \tag{12}$$

It is clearly seen that such a maximizer is bounded at $Y_{(n)}$ and bounded away from zero at $Y_{(1)}$ by assumption (4) because otherwise the weighted log likelihood function (3) becomes $-\infty$, which contradicts the maximization.

We now establish the iterative procedure based on the profile likelihood idea for calculating $(\hat{\beta}_n, \hat{\Lambda}_n)$: (i) for a fixed β, $\hat{\Lambda}_n(\cdot;\beta)$ can be computed iteratively using the iterative convex minorant algorithm described above through updating (8), (9), (10), (11) and the left derivative of the greatest convex minorant of the cumulative sum diagram defined by (0, 0) and the points in (12); (ii) then

β can be updated by solving Equation (5) using the Newton–Raphson algorithm; and (iii) repeat the process until convergence. The initial value of β may be chosen as 0. Simulation shows that the algorithm converges very quickly.

3. ASYMPTOTIC PROPERTIES

We present the asymptotic properties of the estimators with true weights and estimated weights separately because their proofs require different techniques. Both are based on the following regularity conditions.

(A) The parameter space for β, $\mathcal{B} \subset R^d$, is compact, and the true parameter β_0 is an interior point of $\mathcal{B}$.

(B) The cumulative hazard function Λ satisfies $1/M \leq \Lambda \leq M$ on $[\sigma, \tau]$ with $\sigma > 0$ for some positive constant M. The true parameter Λ_0 satisfies $0 < \Lambda_0(\sigma) < \Lambda_0(\tau) < M$ and is continuously differentiable with positive derivative on $[\sigma, \tau]$.

(C) The function $r(\cdot)$ is positive, bounded away from zero, and twice continuously differentiable.

(D) The inspection time Y possesses a Lebesgue density that is continuous and positive on the interval $[\sigma, \tau]$ and vanishes outside this interval, and the joint distribution $F(y, z)$ of (Y, Z) has bounded second-order partial derivative with respect to y.

(E) The covariate vector Z is bounded, and $E[\text{var}(Z|Y)]$ and $E[\text{var}(Z\dot{v}(a^T Z)|Y)]$ are positive definite for all constant vector $a \in R^d$, where $v(\cdot) = \log r(\cdot)$.

(F) There exists a constant ε such that $\pi_{\alpha_0}(\Delta, V) \geq \varepsilon > 0$, where α_0 is the true value of α.

(F') There exists a constant ε such that $\pi_\alpha(\Delta, V) \geq \varepsilon > 0$ for all α in a neighborhood of the true parameter α_0.

Denote the parameter space for Λ defined in (B) by Φ and the parameter space for (β, Λ) by Θ. The above Assumptions (A), (B), (D), and (E) are basically the same as those in van der Vaart (2002) for the full data Cox model with current status data. They are imposed mainly for technical reasons, but also make practical sense. For instance, τ can be viewed as the time of the end of study. Assumption (D), an important condition for asymptotic normality in the complete data case, can be simplified when Y and Z are independent, which reduces to a condition only for the marginal distribution of Y. Assumption (E) ensures the identifiability of the model as well as the positive definiteness of the efficient information matrix for β. For the Cox model, condition $E[\text{var}(Z\dot{v}(a^T Z)|Y)] > 0$ in (E) is redundant. The positivity requirement in (C) may be weakened as in Prentice & Self (1983), but such a requirement is cleaner for the theoretical derivation and can be achieved in the numerical implementation by using, for example, step-halving, that is reducing the search depth of β by half in the Newton–Raphson iteration when the assumption is violated. Assumption (A) and the boundedness of Z in Assumption (E), though not necessary, are helpful in ensuring Assumption (C) for models like $r(t) = 1 + t$. Assumption (F) is for the case of using true weights and Assumption (F'), a stronger condition than Assumption (F), is for the case of using estimated weights, which are commonly assumed for missing data problems. The parameter space of α is unspecified in Assumption (F') for generality. Later in Theorems 3 and 4 we will see that the estimator of α needs to have a root-n rate, hence choosing a parametric model for π_α is a reasonable consideration. This is indeed the case for stratified sampling with finite number of strata for variable V.

3.1. Estimation With True Weights

Let $|\cdot|$ be the Euclidian norm and $\|\Lambda\|_2 = \{\int \Lambda^2(y)dQ_Y(y)\}^{1/2}$ for every $\Lambda \in \Phi$, where $Q_Y(y)$ is the probability measure of the inspection time Y. Define the distance in $\Theta \equiv \mathcal{B} \times \Phi$ as $d((\beta_1, \Lambda_1), (\beta_2, \Lambda_2)) = |\beta_1 - \beta_2| + \|\Lambda_1 - \Lambda_2\|_2$. Given the true weights $w_i = \xi_i/\pi_{\alpha_0}(\Delta_i, V_i)$, we then have the following consistency result with a proof provided in Appendix B.

Theorem 1. *Under Assumptions (A) to (F), we have $\hat\beta_n \to_p \beta_0$ and $\hat\Lambda_n(t) \to_p \Lambda_0(t)$ for every $t \in (\sigma, \tau)$ as $n \to \infty$. Furthermore, we have $|\hat\beta_n - \beta_0| + \|\hat\Lambda_n - \Lambda_0\|_2 = O_p(n^{-1/3})$.*

In fact, the convergence of $(\hat\beta_n, \hat\Lambda_n)$ also holds almost surely, but convergence in probability suffices for our purpose. Note that we only need the pointwise convergence of $\hat\Lambda_n$ in the open interval (σ, τ) to obtain the desirable asymptotic distribution for $\hat\beta_n$, the estimator of our primary parameter of interest. It is natural to see that, as in the complete data case for the Cox model, the overall rate of convergence for the missing data problem for the general relative risk regression is also dominated by $\hat\Lambda_n$ that has a cubic root-n rate. The next theorem shows that the rate of convergence of $\hat\beta_n$ is the usual root-n rate and is asymptotically normal.

When there is no missing data, the efficient score function for β in model (1) can be calculated similarly as in Huang (1996) by the projection method of Bickel et al. (1993). In particular, the usual score function for β is $\dot\ell_1(\beta, \Lambda; X) = \partial l(\beta, \Lambda; X)/\partial\beta = \dot r(\beta^T Z)\Lambda(Y)Q(X)Z$ and the score operator for Λ is $\dot\ell_2(\beta, \Lambda; X)[h] = \partial l(\beta, \Lambda_\eta; X)/\partial\eta = r(\beta^T Z)Q(X)h(Y)$ for every $h \in H \equiv \{h : h = \partial\Lambda_\eta/\partial\eta|_{\eta=0}\}$ with $\Lambda = \Lambda_{\eta=0}$, where

$$Q(X) = \Delta\frac{\exp(-\Lambda(Y)r(\beta^T Z))}{1-\exp(-\Lambda(Y)r(\beta^T Z))} - (1-\Delta).$$

It follows that the efficient score for β has the following form:

$$\begin{aligned}\tilde l(\beta, \Lambda; X) &= \dot\ell_1(\beta, \Lambda; X) - \dot\ell_2(\beta, \Lambda; X)[\mathbf{h}^*]\\ &= \Lambda(Y)Q(X)\left[\dot r(\beta^T Z)Z - r(\beta^T Z)\tfrac{E\{Z\dot r(\beta^T Z)r(\beta^T Z)u(Y,Z;\beta,\Lambda)|Y\}}{E\{r^2(\beta^T Z)u(Y,Z;\beta,\Lambda)|Y\}}\right],\end{aligned} \tag{13}$$

where

$$\begin{aligned} u(Y, Z; \beta, \Lambda) &= \frac{\exp(-\Lambda(Y)r(\beta^T Z))}{1-\exp(-\Lambda(Y)r(\beta^T Z))}, \quad \text{and}\\ \mathbf{h}^*(y) &= \Lambda(y)\frac{E\{Z\dot r(\beta^T Z)r(\beta^T Z)u(Y, Z; \beta, \Lambda)|Y = y\}}{E\{r^2(\beta^T Z)u(Y, Z; \beta, \Lambda)|Y = y\}}\end{aligned} \tag{14}$$

is the least favourable direction. The information matrix for β is then given by

$$I(\beta) = E\left\{\tilde l(\beta, \Lambda; X)^{\otimes 2}\right\}, \tag{15}$$

where $v^{\otimes 2} = vv^T$ for a vector v. When $r(\cdot) = \exp(\cdot)$, these results reduce to that of Huang (1996), Murphy & van der Vaart (2000), and van der Vaart (2002).

The following theorem states the asymptotic normality for the weighted likelihood estimator $\hat\beta_n$ obtained by using true weights. We can see that the asymptotic variance matrix is the complete data asymptotic variance matrix, the inverse of (15) at β_0, plus an additional non-negative definite matrix that reflects the loss of efficiency due to missing data.

Theorem 2. *Under Assumptions (A) to (F), we have*

$$\sqrt{n}\left(\hat\beta_n - \beta_0\right) = I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i\tilde l(\beta_0, \Lambda_0; X_i) + o_p(1) \to_d N(0, \Sigma)$$

as $n \to \infty$, where $\Sigma = I^{-1}(\beta_0) + I^{-1}(\beta_0)DI^{-1}(\beta_0)$, and

$$D = E\left[\frac{1-\pi_{\alpha_0}(\Delta, V)}{\pi_{\alpha_0}(\Delta, V)}\left\{\tilde l(\beta_0, \Lambda_0; X)\right\}^{\otimes 2}\right].$$

3.2. Estimation With Estimated Weights

In this subsection we denote the weight by $w(\alpha)$, where $\alpha = (\alpha_1, \ldots, \alpha_J)^T$ with true value $\alpha_0 = (\alpha_{01}, \ldots, \alpha_{0J})^T$. No matter α_0 is known or not, we can replace it by a good estimator $\hat{\alpha}_n = (\hat{\alpha}_{n1}, \ldots, \hat{\alpha}_{nJ})^T$, then use the estimated weight $w(\hat{\alpha}_n)$ in the weighted likelihood function. Let

$$(\tilde{\beta}_n, \tilde{\Lambda}_n) = \operatorname{argmax} \sum_{i=1}^{n} w_i(\hat{\alpha}_n) l(\beta, \Lambda; X_i)$$

be the weighted likelihood estimator of (β_0, Λ_0) obtained by using estimated weights.

When nuisance parameters can be estimated at the root-n rate, the efficiency gain of the estimator $\tilde{\beta}_n$ comparing to $\hat{\beta}_n$ that is obtained using true weights has been discussed by many authors, for example, Pierce (1982), Robins, Rotnitzky, & Zhao (1994), Breslow & Wellner (2007), and Li et al. (2008), among many others. In particular, a heuristic argument was provided by Robins et al. (1994) in their Discussion Section. It turns out that for the current setting in which the infinite-dimensional nuisance parameter can only be estimated at a slower than root-n rate, such an efficiency gain for the estimation of the parameter of interest also holds under mild conditions (see Theorem 4).

We first give the results of consistency and rate of convergence for $\tilde{\beta}_n$ in the following Theorem 3.

Theorem 3. *Suppose $\hat{\alpha}_n \to_p \alpha_0$ and $w(\alpha)$ is differentiable with uniformly bounded first-order derivative $\dot{w}(\alpha)$ in a neighborhood of α_0. Then under Assumptions (A) to (E) and (F'), we have $\tilde{\beta}_n \to_p \beta_0$ and $\tilde{\Lambda}_n(t) \to_p \Lambda_0(t)$ for every $t \in (\sigma, \tau)$. If further assume that* $\sup_n E\sqrt{n}|\hat{\alpha}_n - \alpha_0| < \infty$ *and $w(\alpha)$ is twice differentiable with uniformly bounded second-order derivative $\ddot{w}(\alpha)$ in a neighborhood of α_0, then $|\tilde{\beta}_n - \beta_0| + \|\tilde{\Lambda}_n - \Lambda_0\|_2 = O_p(n^{-1/3})$.*

The uniform boundedness of $\dot{w}(\alpha)$ and $\ddot{w}(\alpha)$ is not too restrictive. For example, for a case-cohort design with a stratified Bernoulli sampled subcohort, we have $\pi_\alpha(\Delta, V) = \Delta + (1-\Delta)\sum_{j=1}^{J} p_j I_{(V \in \mathcal{V}_j)}$, and the above conditions are satisfied as long as all the stratified selection probabilities p_j's are bounded away from 0. The same is true for a two-phase design in which the second stage sample is selected by a stratified Bernoulli sampling. More generally, if $\pi_\alpha(\Delta, V)$ follows a logistic model, say, logit $\pi_\alpha(\Delta, V) = \alpha_0 + \alpha_1^T V + \alpha_2 \Delta$, then the conditions are still satisfied given that V is bounded. The boundedness of $\sup_n E\sqrt{n}|\hat{\alpha}_n - \alpha_0|$ is a little more restrictive. The asymptotic normality of $\sqrt{n}(\hat{\alpha}_n - \alpha_0)$ is neither sufficient nor necessary for this to hold, while the condition that $E\sqrt{n}|\hat{\alpha}_n - \alpha_0|$ converges to a finite limit as $n \to \infty$ is stronger than necessary. Nevertheless, in a case-cohort design, or a more general two-phase stratified sampling design, $\hat{p}_j$ is the proportion of subjects selected from stratum j, $1 \le j \le J$, and it is easy to show that $E\sqrt{n}|\hat{p}_j - p_{0j}|$ converges to a finite limit as $n \to \infty$.

The following theorem shows the asymptotic normality of $\tilde{\beta}_n$ as well as the efficiency gain of $\tilde{\beta}_n$ comparing to $\hat{\beta}_n$, which will be proved in Appendix B by applying the general theorem introduced in Appendix A.

Theorem 4. *Under the same conditions in Theorem 3, we have*

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) = I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i(\alpha_0)\tilde{l}(\beta_0, \Lambda_0; X_i) - C\sqrt{n}(\hat{\alpha}_n - \alpha_0) + o_p(1)$$

as $n \to \infty$, where $C = I^{-1}(\beta_0)P\{\tilde{l}(\beta_0, \Lambda_0; X)\dot{w}^T(\alpha_0)\}$. Furthermore, if $\hat{\alpha}_n$ is asymptotically efficient with the following asymptotic representation:

$$\hat{\alpha}_n = \alpha_0 + \frac{1}{n}\sum_{i=1}^{n} \ell^{\alpha}(X_i) + o_p\left(n^{-1/2}\right),$$

then we have

$$\sqrt{n}(\tilde{\beta}_n - \beta_0) \to_d N(0, \Sigma - C\Sigma_{\alpha}C^T),$$

where Σ was defined in Theorem 2 and $\Sigma_{\alpha} = E(\ell^{\alpha\otimes 2})$.

Note that the function ℓ^{α} in the above asymptotic representation of $\hat{\alpha}_n$ has zero mean and is called the (efficient) influence function of $\hat{\alpha}_n$. We refer to Bickel et al. (1993) for a thorough discussion on influence functions.

4. VARIANCE ESTIMATION

As discussed in Huang (1996), directly applying the asymptotic variance expressions in Theorems 2 and 4 for the variance estimation requires smoothing. The weighted bootstrap with i.i.d. weights, however, turns out to be an effective and robust approach in variance estimation for the weighted likelihood estimator with either true weights or estimated weights without applying any smoothing technique. See Ma & Kosorok (2005) for details of using the weighted bootstrap method for the general M-estimation in semiparametric models.

Firstly consider the case in which true weights are used. Suppose that $u_1, \ldots, u_n$ are n i.i.d. non-negative and bounded random weights, independent of $X_1, \ldots, X_n$ and $w_1, \ldots, w_n$, and satisfying $E(u_i) = 1$ and $\text{var}(u_i) = \delta_0 < \infty$ for a constant δ_0. Denote the estimator of β obtained by maximizing the objective function $\sum_{i=1}^{n} u_i w_i l(\beta, \Lambda; X_i)$ by $\hat{\beta}_n^*$. Randomly generate $(u_1, \ldots, u_n)$ repeatedly, say, B times, and obtain corresponding $\hat{\beta}_n^*$ that are denoted by $\hat{\beta}_{n1}^*, \ldots, \hat{\beta}_{nB}^*$. A variance estimator of $\hat{\beta}_n$ is then obtained from the empirical variance of $\hat{\beta}_{n1}^*, \ldots, \hat{\beta}_{nB}^*$ rescaled by δ_0. Analogous to the case in Ma & Kosorok (2005), this weighted bootstrap estimation of variance can be justified in the following way.

Since u is bounded with mean 1 and independent of X_i's and w_i's, we have $E\{uwl(\beta, \Lambda; X)\} = E\{wl(\beta, \Lambda; X)\}$. By Theorem 2 we have

$$\sqrt{n}(\hat{\beta}_n^* - \beta_0) = I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i w_i \tilde{l}(\beta_0, \Lambda_0; X_i) + o_p(1).$$

Hence

$$\sqrt{n}(\hat{\beta}_n^* - \hat{\beta}_n) = I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} (u_i - 1) w_i \tilde{l}(\beta_0, \Lambda_0; X_i) + o_p(1).$$

Then by Theorem 2 of Ma & Kosorok (2005) we know that, conditional on data $(X_1, w_1), \ldots, (X_n, w_n)$, $(n/\delta_0)^{1/2}(\hat{\beta}_n^* - \hat{\beta}_n)$ has the same asymptotic distribution as that of $\sqrt{n}(\hat{\beta}_n - \beta_0)$ unconditionally.

When estimated weights are used in the weighted likelihood, additional care needs to be taken to make the weighted bootstrap work. Specifically, in addition to multiplying each term in the original estimating equation by a bootstrap weight u_i, the parameter α needs to be estimated again by the weighted bootstrap using the same set of weights. We can show, in a way similar to the above using the true weights, this procedure yields valid variance estimates. Since for the original $\tilde{\beta}_n$ and $\hat{\alpha}_n$ we have

$$\sqrt{n}(\tilde{\beta}_n-\beta_0) = I^{-1}(\beta_0)\frac{1}{\sqrt{n}}\sum_{i=1}^{n} w_i\tilde{l}(\beta_0, \Lambda_0; X_i)-C\frac{1}{\sqrt{n}}\sum_{i=1}^{n} l^{\alpha}(X_i) + o_p(1)$$

and

$$\sqrt{n}(\hat{\alpha}_n-\alpha_0) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} l^{\alpha}(X_i) + o_p(1)$$

by Theorem 4, and for the weighted bootstrap estimate of $\hat{\alpha}_n^*$ we have

$$\sqrt{n}\left(\hat{\alpha}_n^*-\alpha_0\right) = \frac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i l^{\alpha}(X_i) + o_p(1),$$

we obtain for the weighted bootstrap estimate $\tilde{\beta}_n^*$ that

$$\begin{aligned}\sqrt{n}(\tilde{\beta}_n^*-\beta_0) &= I^{-1}(\beta_0)\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i w_i\tilde{l}(\beta_0, \Lambda_0; X_i)-C\sqrt{n}(\hat{\alpha}_n^*-\alpha_0) + o_p(1)\\ &= I^{-1}(\beta_0)\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i w_i\tilde{l}(\beta_0, \Lambda_0; X_i)-C\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} u_i l^{\alpha}(X_i) + o_p(1).\end{aligned}$$

It follows that

$$\begin{aligned}\sqrt{n}\left(\tilde{\beta}_n^*-\tilde{\beta}_n\right) &= I^{-1}(\beta_0)\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} (u_i-1)w_i\tilde{l}(\beta_0, \Lambda_0; X_i)\\ &\quad - C\tfrac{1}{\sqrt{n}}\sum_{i=1}^{n} (u_i-1)l^{\alpha}(X_i) + o_p(1).\end{aligned}$$

Therefore, conditional on $(X_1, w_1), \ldots, (X_n, w_n)$, $(n/\delta_0)^{1/2}(\tilde{\beta}_n^*-\tilde{\beta}_n)$ has the same asymptotic distribution as $\sqrt{n}(\tilde{\beta}_n-\beta_0)$. Note that updating the estimator of α in the bootstrap step is required. Otherwise the weighted bootstrap procedure is estimating the variance of the weighted likelihood estimator with true weights, which is clearly not desirable.

5. NUMERICAL RESULTS

5.1. Simulations

A simulation study is conducted to explore the performance of the proposed weighted likelihood estimators. We assume the unobserved event time T follows (i) a proportional hazards model given covariate Z with a constant baseline hazard function $\lambda(t) \equiv c$, which implies that the failure time has an exponential distribution or (ii) a linear relative risk model with a constant baseline hazard function. The inspection time Y is assumed to be uniformly distributed in the interval between 0.5 and 8.5. The covariate Z has two components Z_1 and Z_2, where $Z_1 \sim N(0, 1)$ truncated at -3 from left and 3 from right, and Z_2 is binary with $Pr(Z_2 = 0) = Pr(Z_2 = 1) = 0.5$. The true parameter for β is $\beta_0 = (1, -1)^T$ for the Cox model and $\beta_0 = (0.2, -0.2)^T$ in the linear relative risk model. We consider two different scenarios. In scenario 1, we set $n = 500$ and take $c = 0.03$ for the Cox model and $c = 0.06$ for the linear relative risk model; In scenario 2, we set $n = 3000$ and take $c = 0.01$ for the Cox model and $c = 0.02$ for the linear relative risk model. We first generate n i.i.d. samples of (Δ, Y, Z) and then generate missing covariates. The missing covariates are generated via a case-cohort sampling method. We assume that Z_1 can be missing while Z_2 is always observed. The probability of missing Z_1 is 0 for a subject with a failure event, and depends on an auxiliary variable V for a censored subject. The auxiliary variable V is associated with the covariates of interest in the following way: $V = 1$ when $Z_1 < 1$ and $Z_2 = 0$, $V = 2$ when $Z_1 < 1$ and $Z_2 = 1$, $V = 3$ when $Z_1 \geq 1$ and $Z_2 = 0$, and $V = 4$ when $Z_1 \geq 1$ and $Z_2 = 1$. When $n = 500$, the probability of observing covariate Z_1 is $P = 0.2$ if $V = 1$ or 2, and $P = 0.7$ if $V = 3$ or 4. When $n = 3{,}000$, $P = 0.05$ if $V = 1$ or 2, and $P = 0.15$ if $V = 3$ or 4. Under these circumstances, when sample size $n = 500$, for the Cox model there are about 170 subjects with covariates fully observed in which about 100 are failures, for the linear relative risk model there are about 200 subjects with covariates fully observed in which about 100 are failures, and $P(T > 8.5) \approx 0.65$ for both models; and when $n = 3{,}000$, for the Cox model there are about 400 subjects with fully observed covariates in which about 250 are failures, for the linear relative risk model there are about 500 subjects with fully observed covariates in which about 250 are failures, and $P(T > 8.5) \approx 0.75$ for both models. The setting for $n = 3{,}000$ here mimics the setting for the HIV case-cohort study in the next subsection.

We then calculate the weighted likelihood estimator $(\hat{\beta}_n, \hat{\Lambda}_n)$ using the iterative algorithm given in The Weighted Likelihood Estimator Section for each generated data set. We choose (0, 0) as the initial value of $\hat{\beta}_n$, and then iterate between $\hat{\beta}_n$ and $\hat{\Lambda}_n$ until convergence. The same procedure is executed to obtain $(\tilde{\beta}_n, \tilde{\Lambda}_n)$, where the estimated weight for each subject with Z_1 observed is the inverse of sample fraction within corresponding stratum determined by (Δ, V). For the linear relative risk model, we use step-halving in updating β to ensure positivity of the risk function. We run 500 replications for the simulation, and then obtain point estimates and biases of the estimators of β_0. Variance estimates are obtained by the weighted bootstrap procedure. To apply the weighted bootstrap method, we generate independent weight u from a uniform distribution on (0, 2), and use 100 bootstrap samples to estimate variance for each simulated data set. We also provide results for the nonparametric maximum likelihood estimator of β when covariates are fully observed (full data MLE), which are calculated by setting the weights for all subjects to be 1. The numerical calculation is implemented in R.

Biases, sample averages of estimated variances, empirical variances, and coverage proportions of 95% confidence intervals for the slope estimators of Z_1 and Z_2 are presented in Table 1. The biases are reasonably small across the board, particularly for the larger sample size. The variance estimates are very close to corresponding empirical variances and yield good coverage proportions. Comparing empirical variances of the weighted likelihood estimates with true weights and those with estimated weights, the efficiency gain of the latter is clearly seen, which supports the theoretical result given in Theorem 4. Plots in Figure 1 are the averages of estimated baseline cumulative hazard functions over 500 simulation replications when sample size is 500. We can see that the average curves using true weights and estimated weights are barely distinguishable. Both estimates have little bias except towards the end of study, a phenomenon also observed in Zhang, Hua, & Huang (2010). Note that the number of fully observed subjects is about 180 when $n = 500$ and both weighted estimates converge at a slow cubic root-n rate. The relative bias reduces about 50% when the sample size increases to $n = 3{,}000$ with about 400 fully observed subjects, and almost disappears when there is no missing data for all $n = 3{,}000$ subjects (results not shown).

5.2. A Case-Cohort Study From an HIV Vaccine Trial

We illustrate our method here by analyzing the case-cohort data collected from one of the largest phase 3 HIV-1 vaccine efficacy trials in the world (Flynn et al., 2005; Gilbert et al., 2005). The trial demonstrated lack of efficacy of the vaccine, but Gilbert et al. (2005) undertook a secondary objective, which was to determine whether antibody responses are correlated with the incidence

TABLE 1: Summary statistics of simulations.

	Full data MLE		True weights		Estimated weights	
Method parameter	β_1	β_2	β_1	β_2	β_1	β_2
Cox Model: Scenario 1						
Bias	−0.022	0.016	−0.038	0.075	−0.040	0.074
Bootstrap Std	0.141	0.145	0.182	0.187	0.167	0.176
Empirical Std	0.145	0.148	0.184	0.195	0.170	0.179
Coverage proportion	0.946	0.946	0.947	0.927	0.950	0.943
Cox Model: Scenario 2						
Bias	0.013	−0.020	−0.013	0.033	−0.010	0.031
Bootstrap Std	0.071	0.084	0.134	0.134	0.110	0.116
Empirical Std	0.077	0.084	0.138	0.138	0.113	0.120
Coverage proportion	0.940	0.945	0.953	0.940	0.948	0.945
Linear Risk Model: Scenario 1						
Bias	−0.007	0.001	0.021	0.008	0.017	0.006
Bootstrap Std	0.170	0.089	0.224	0.138	0.212	0.101
Empirical Std	0.164	0.095	0.212	0.130	0.200	0.105
Coverage Proportion	0.946	0.930	0.937	0.930	0.935	0.931
Linear Risk Model: Scenario 2						
Bias	−0.001	0.005	0.004	−0.011	−0.007	−0.008
Bootstrap Std	0.100	0.060	0.235	0.170	0.202	0.114
Empirical Std	0.105	0.063	0.251	0.164	0.217	0.100
Coverage proportion	0.941	0.938	0.930	0.952	0.924	0.956

of HIV-1 infection among vaccine recipients. The trial was designed to have multiple visits and either vaccine or placebo was administered at each visit. For simplicity, we only consider the infection status at the last visit and thus have the current status data to work with. The approach of analyzing interval censored data with multiple random inspection times is under investigation. The original trial consists of 5,095 men and 308 women who received the study vaccine or placebo at a 2:1 ratio. Each study participant was followed up to 36 months. Gilbert et al. (2005) designed a case-cohort study that consisted of all 241 infected subjects and 167, a fraction of 5%, uninfected subjects selected via independent Bernoulli sampling, all being selected from vaccine recipients. This is a classical case-cohort design without covariate stratification. They found that the peak antibody levels reached a high level at month 6.5 (after the second vaccine shot) and became relatively stable afterwards. We consider the only functional assay, the MN neutralization titer ($\min = 1.48$, $\text{median} = 2.83$, $\max = 5.07$), among all antibody responses and use its peak level at month 6.5 (hence infections prior month 6.5 are excluded) as the covariate of interest in our analysis. This antibody in principle should be most relevant for HIV protection. The Cox proportional hazards model is considered and a cubic-root power transformation of the antibody peak level is used to achieve a better linear effect in the Cox model. Several demographic variables are also considered in the Cox proportional hazards model, but only the baseline behavioral risk score is significant. Since only the sample fraction of 5%, the most efficient estimator of the true

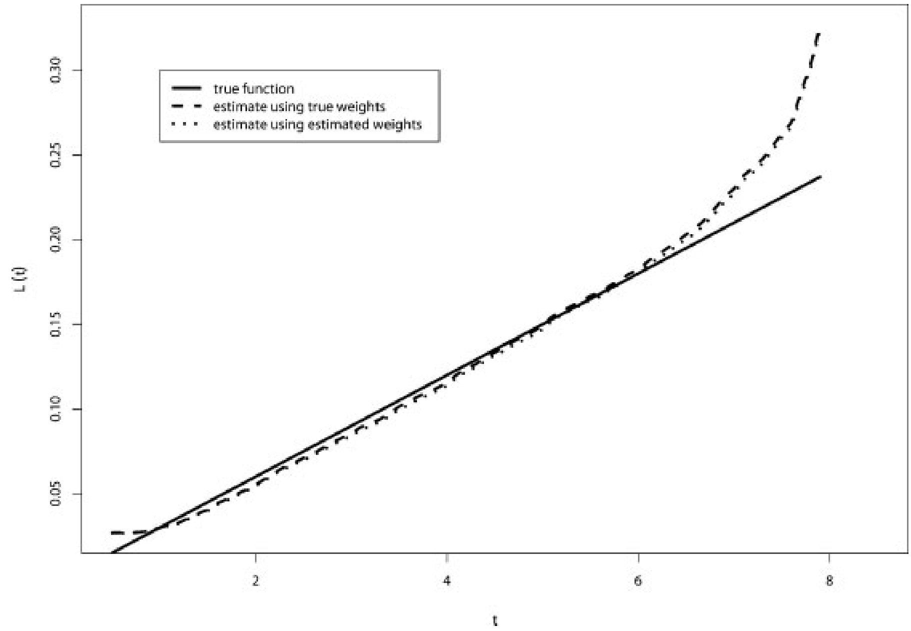

FIGURE 1: Average estimates of the baseline cumulative hazard function over 500 simulation replications when sample size $n = 500$.

TABLE 2: Estimates of log hazard ratios for MN neutralizing titer (MN) and the baseline behavioral risk score.

Variable	MN	Low risk	Medium risk	High risk
Estimate	−0.654	0	0.898	2.385
Std. Error	0.324	—	0.249	0.542
P-value	0.043	—	<0.001	<0.001

Low risk (reference group): the group with risk scores equal to 0.

Medium risk: the group with risk scores from 1 to 3.

High risk: the group with risk scores greater than 3.

selection probability, for uninfected subjects was provided by Gilbert et al. (2005), we use the weighted likelihood method with estimated weights in our analysis that should yield more efficient regression parameter estimation than using true weights (Theorem 4). The final result is given in Table 2. We can see that the antibody MN neutralization titer has a protection effect against HIV infection, which is consistent with the finding in Gilbert et al. (2005) where an analysis of approximated right censored data was conducted.

6. DISCUSSION

The proposed weighted likelihood method can be applied to stratified sampling designs when complete data are selected by an i.i.d. Bernoulli sampling that results in an i.i.d. structure of the data. An alternative practical sampling approach is sampling without replacement, wherein the number of sampled subjects in each stratum is fixed. Such a sampling design destroys the i.i.d. data structure. Breslow & Wellner (2007) considered this type of designs for semiparametric models in which the infinite dimensional nuisance parameter can be estimated at a root-n rate, and provided proofs of asymptotic properties based on the weighted bootstrap empirical process theory of Præstgaard & Wellner (1993). An interesting work that is undergoing is to extend the work of Breslow & Wellner (2007) to two-phase designs with current status data where the second phase data are selected by sampling without replacement, in which the baseline cumulative hazard function should still only be estimable at a cubic root-n rate.

ACKNOWLEDGEMENTS

This work was supported in part by NSF grants DMS-0706700 and DMS-1007590. The authors thank Peter Gilbert and VaxGen, Inc. for providing the HIV vaccine trial data, and Xuwen Lu for helpful discussions on the weighted bootstrap method.

APPENDIX A: AN ASYMPTOTIC NORMALITY THEOREM FOR SEMIPARAMETRIC M-ESTIMATION

We extend Theorem 6.1 of Wellner & Zhang (2007) by replacing one of the nuisance parameters by its estimator in the objective function that will be maximized with respect to all other parameters. Such an extension is crucial in handling the missing data problem when weights are estimated, and can be useful in proving asymptotic normality for a general semiparametric missing data problem when the missing probability is estimated from observed data. For simplicity of notation, we adopt the empirical process notation of van der Vaart and Wellner throughout the Appendices by denoting Pf as the integral of f with respect to the probability measure P, $\mathbb{P}_n f$ as the integral of f with respect to the empirical measure $\mathbb{P}_n$, which is the sample average of f for i.i.d. data, and $\mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f$.

Given i.i.d. observations $X_1, \cdots, X_n$, suppose that the estimates $(\tilde\beta_n,\ \tilde\Lambda_n)$ of unknown parameters $(\beta,\ \Lambda)$ are set to be the maximizer of the objective function $\mathbb{P}_n m(\beta, \Lambda, \hat\alpha_n; X)$, where $\hat\alpha_n$ is an estimator of the true parameter α_0, $\beta \in R^d$, and $\Lambda \in \Phi$, an infinite dimensional Banach space. Here we assume α_0 to be finite dimensional, though it can be more general. Suppose that Λ_η is a parametric submodel in Φ passing through Λ, that is, $\Lambda_\eta \in \Phi$ and $\Lambda_{\eta=0} = \Lambda$. Let $H = \{h : h = \partial\Lambda_\eta/\partial\eta|_{\eta=0}\}$ be the collection of all directions to approach Λ. Let $\dot m_1(\beta, \Lambda, \alpha; X) = \partial m(\beta, \Lambda, \alpha; X)/\partial\beta$, $\dot m_2(\beta, \Lambda, \alpha; X)[h] = \partial m(\beta, \Lambda_\eta, \alpha; X)/\partial\eta$ along the direction of h, and $\dot m_3(\beta, \Lambda, \alpha; X) = \partial m(\beta, \Lambda, \alpha; X)/\partial\alpha$. Let $\ddot m_{ij}$ be the second order derivatives of m with respect to corresponding arguments defined in a similar way, $i, j \in \{1, 2, 3\}$.

The following conditions are mostly parallel to those in Theorem 6.1 of Wellner & Zhang (2007), but here they are adapted to accommodate a more general setting.

A1. $|\hat\alpha_n - \alpha_0| = o_p(1)$, $|\tilde\beta_n - \beta_0| = o_p(1)$, and $\|\tilde\Lambda_n - \Lambda_0\| = O_p(n^{-\gamma})$ for some $\gamma > 0$ and some norm $\|\cdot\|$.

A2. There exists an $\mathbf{h}^* = (h_1^*, \cdots, h_d^*)^T$, where $h_j^* \in L_2(P)$, $j = 1, 2, \cdots, d$, such that

$$P\{\ddot m_{12}(\beta_0, \Lambda_0, \alpha_0; X)[h] - \ddot m_{22}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*, h]\} = 0,$$

for all $h \in H$. Moreover, the matrix

$$A = -P\{\ddot m_{11}(\beta_0, \Lambda_0, \alpha_0; X) - \ddot m_{21}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*]\}$$

is non-singular.

A3. $P\dot m_1(\beta_0, \Lambda_0, \alpha_0; X) = 0$ and $P\dot m_2(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*] = 0$.

A4. The estimator $(\tilde\beta_n, \tilde\Lambda_n)$ satisfies

$$\mathbb{P}_n \dot m_1(\tilde\beta_n, \tilde\Lambda_n, \hat\alpha_n; X) = o_p(n^{-1/2}) \quad \text{and} \quad \mathbb{P}_n \dot m_2(\tilde\beta_n, \tilde\Lambda_n, \hat\alpha_n; X)[\mathbf{h}^*] = o_p(n^{-1/2}).$$

A5. For any $\delta_n \downarrow 0$ and $C > 0$, let

$$\Theta_n = \{(\beta, \Lambda, \alpha) : |(\beta^T, \alpha^T) - (\beta_0^T, \alpha_0^T)| \le \delta_n, \|\Lambda - \Lambda_0\|_2 \le Cn^{-\gamma}\}.$$

We have

$$\sup_{(\beta,\Lambda,\alpha)\in\Theta_n} |\mathbb{G}_n\{\dot m_1(\beta, \Lambda, \alpha; X) - \dot m_1(\beta_0, \Lambda_0, \alpha_0; X)\}| = o_p(1),$$

and

$$\sup_{(\beta,\Lambda,\alpha)\in\Theta_n} \left|\mathbb{G}_n\left\{\dot m_2(\beta, \Lambda, \alpha; X)[\mathbf{h}^*] - \dot m_2(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*]\right\}\right| = o_p(1).$$

A6. For some $\mu > 1$ satisfying $\mu\gamma > 1/2$, and for $(\beta, \Lambda, \alpha) \in \Theta_n$,

$$\begin{aligned}&|P\{\dot m_1(\beta, \Lambda, \alpha; X) - \dot m_1(\beta_0, \Lambda_0, \alpha_0; X) - \ddot m_{11}(\beta_0, \Lambda_0, \alpha_0; X)(\beta - \beta_0)\\&\quad - \ddot m_{12}(\beta_0, \Lambda_0, \alpha_0; X)[\Lambda - \Lambda_0] - \ddot m_{13}(\beta_0, \Lambda_0, \alpha_0; X)(\alpha - \alpha_0)\}|\\&= o(|\beta - \beta_0|) + o(|\alpha - \alpha_0|) + O(\|\Lambda - \Lambda_0\|^{\mu}),\end{aligned}$$

and

$$\begin{aligned}&|P\{\dot m_2(\beta, \Lambda, \alpha; X)[\mathbf{h}^*] - \dot m_2(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*] - \ddot m_{21}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*](\beta - \beta_0)\\&\quad - \ddot m_{22}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*, \Lambda - \Lambda_0] - \ddot m_{23}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*](\alpha - \alpha_0)\}|\\&= o(|\beta - \beta_0|) + o(|\alpha - \alpha_0|) + O(\|\Lambda - \Lambda_0\|^{\mu}).\end{aligned}$$

Theorem A.1. *Suppose that Conditions A1 to A6 hold. Then we have*

$$\sqrt{n}(\tilde\beta_n - \beta_0) = A^{-1}\sqrt{n}\mathbb{P}_n \dot m^*(\beta_0, \Lambda_0, \alpha_0; X) - C\sqrt{n}(\hat\alpha_n - \alpha_0) + o_{p^*}(1), \tag{A.1}$$

where

$$\dot m^*(\beta_0, \Lambda_0, \alpha_0; X) = \dot m_1(\beta_0, \Lambda_0, \alpha_0; X) - \dot m_2(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*],$$

and

$$C = A^{-1}P\{\ddot m_{13}(\beta_0, \Lambda_0, \alpha_0; X) - \ddot m_{23}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*]\}.$$

If $\sqrt{n}(\hat\alpha_n - \alpha_0)$ is asymptotically normal with influence function ℓ^α, then $\sqrt{n}(\tilde\beta_n - \beta_0)$ is asymptotically normal. Furthermore, if $\hat\alpha_n$ is asymptotically efficient, then $\sqrt{n}(\tilde\beta_n - \beta_0) \to_d N(0, \Omega)$ with

$$\Omega = A^{-1}\left\{P\dot m^*(\beta_0, \Lambda_0, \alpha_0; X)^{\otimes 2}\right\}(A^{-1})^T - C(P\ell^{\alpha\otimes 2})C^T.$$

Proof. By A1, A3 and A5, we have

$$\mathbb{P}_n \dot m_1(\tilde\beta_n, \tilde\Lambda_n, \hat\alpha_n; X) - P\dot m_1(\tilde\beta_n, \tilde\Lambda_n, \hat\alpha_n; X) - \mathbb{P}_n \dot m_1(\beta_0, \Lambda_0, \alpha_0; X) = o_p(n^{-1/2}).$$

In view of A4, this reduces to

$$P\dot m_1(\tilde\beta_n, \tilde\Lambda_n, \hat\alpha_n; X) + \mathbb{P}_n \dot m_1(\beta_0, \Lambda_0, \alpha_0; X) = o_p(n^{-1/2}).$$

Then by A6, it follows that

$$\begin{aligned} &P\ddot{m}_{11}(\beta_0, \Lambda_0, \alpha_0; X)(\tilde{\beta}_n - \beta_0) + P\ddot{m}_{12}(\beta_0, \Lambda_0, \alpha_0; X)[\tilde{\Lambda}_n - \Lambda_0] \\ &\quad + P\ddot{m}_{13}(\beta_0, \Lambda_0, \alpha_0; X)(\hat{\alpha}_n - \alpha_0) + \mathbb{P}_n \dot{m}_1(\beta_0, \Lambda_0, \alpha_0; X) \\ &= o(|\tilde{\beta}_n - \beta_0|) + o(|\hat{\alpha}_n - \alpha_0|) + O(\|\tilde{\Lambda}_n - \Lambda_0\|^2) \\ &= o_p(n^{-1/2}), \end{aligned} \tag{A.2}$$

In a similar way, we obtain

$$P\dot{m}_2(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X)[\mathbf{h}^*] + \mathbb{P}_n \dot{m}_2(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*] = o_p(n^{-1/2}),$$

and then

$$\begin{aligned} &P\ddot{m}_{21}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*](\tilde{\beta}_n - \beta_0) + P\ddot{m}_{22}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*, \tilde{\Lambda}_n - \Lambda_0] \\ &\quad + P\ddot{m}_{23}(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*](\hat{\alpha}_n - \alpha_0) + \mathbb{P}_n \dot{m}_2(\beta_0, \Lambda_0, \alpha_0; X)[\mathbf{h}^*] \\ &= o(|\tilde{\beta}_n - \beta_0|) + o(|\hat{\alpha}_n - \alpha_0|) + O(\|\tilde{\Lambda}_n - \Lambda_0\|^2) \\ &= o_p(n^{-1/2}). \end{aligned} \tag{A.3}$$

Subtracting (A.3) from (A.2) and rearranging terms, by A2 we obtain (A.1). When $\sqrt{n}(\hat{\alpha}_n - \alpha_0)$ is asymptotically normal with influence function ℓ^{α}, the right hand side of the above equation converges to a zero mean normal random variable by the classical central limit theorem. Furthermore, when $\hat{\alpha}_n$ is efficient, $\sqrt{n}(\tilde{\beta}_n - \beta_0) \rightarrow_d N(0, \Omega)$ follows from (A.1) and the result in Pierce (1982), with Ω being stated in the theorem.

APPENDIX B: PROOFS OF THEORETICAL RESULTS

Proof of Theorem 1. Following van der Vaart (2002), we introduce the following functions $\ell(\beta, \Lambda; X) = \log\{(p_{\beta,\Lambda} + p_0)/2\}$ and $m(\beta, \Lambda; X) = w\ell(\beta, \Lambda; X)$, where $p_0 = p_{\beta_0,\Lambda_0}$. Although $\mathbb{P}_n m(\beta, \Lambda; X)$ is not maximized at $(\hat{\beta}_n, \hat{\Lambda}_n)$, it is always true that $\mathbb{P}_n m(\hat{\beta}_n, \hat{\Lambda}_n; X) \geq \mathbb{P}_n m(\beta_0, \Lambda_0; X)$. Only this less restrictive condition is needed in Theorem 5.8 in van der Vaart (2002). Note that, under our assumptions, p_0 is bounded and bounded away from 0, so it follows that $m(\beta, \Lambda; X)$ is uniformly bounded. Then by Theorem 5.8 and Lemma 5.9 in van der Vaart (2002), to prove the consistency of $(\hat{\beta}_n, \hat{\Lambda}_n)$, it suffices to show that the parameter space for (β, Λ) is compact, the map $(\beta, \Lambda) \mapsto p_{\beta,\Lambda}(x)$ is continuous for every x, and the map $(\beta, \Lambda) \mapsto Pm(\beta, \Lambda; X)$ achieves a unique maximum at (β_0, Λ_0).

The compactness of the parameter space $\mathcal{B}$ of β is from Assumption (A). By the theorem on page 239 of Billingsley (1999), the parameter space Φ of Λ is compact if Φ is closed and for each sequence $\{\Lambda_n, n \geq 1\}$ in Φ, there exists a subsequence $\{\Lambda_{n'}\}$ and some $\Lambda^* \in \Phi$ such that $\|\Lambda_{n'} - \Lambda^*\|_2 \rightarrow 0$ as $n' \rightarrow \infty$. The closeness of Φ is clearly seen. Now for any sequence $\{\Lambda_n, n \geq 1\}$ in Φ, by the same diagonal argument used in proving Helly's selection theorem (e.g., Billingsley, 1995, p. 336), there exists a subsequence $\{\Lambda_{n'}\}$ and some Λ^* such that $|\Lambda_{n'}(y) - \Lambda^*(y)| \rightarrow 0$ for every continuity point of Λ^*. But this implies, by the dominated convergence theorem, that $\|\Lambda_{n'} - \Lambda^*\|_2 \rightarrow 0$ since the density of Y is bounded. Therefore, Φ is compact. The continuity of the map $(\beta, \Lambda) \mapsto p_{\beta,\Lambda}(x)$ for every x is clearly seen from Equation (1) and Assumption (C).

We now show that the map $(\beta, \Lambda) \mapsto Pm(\beta, \Lambda; X)$ achieves a unique maximum at (β_0, Λ_0). By the fact that $E(w|\Delta, V) = 1$, we have $P\{m(\beta, \Lambda; X) - m(\beta_0, \Lambda_0; X)\} = P\{\ell(\beta, \Lambda; X) - \ell(\beta_0, \Lambda_0; X)\}$ that is negative Kullback–Leibler divergence and hence is always less than or equal to 0. It is 0 if and only if $p_{\beta,\Lambda} = p_0$ with probability 1, or equivalently, $r(\beta^T Z)\Lambda(Y) = r(\beta_0^T Z)\Lambda_0(Y)$ with probability 1. This is equivalent to $v(\beta^T Z) - v(\beta_0^T Z) = -\log\Lambda(Y) + \log\Lambda_0(Y)$ with probability 1, where $v = \log r$. By the Taylor expansion, this can be rewritten as $\dot{v}(a^T Z)(\beta - \beta_0)^T Z = -\log\Lambda(Y) + \log\Lambda_0(Y)$ for some vector a between β and β_0.

This yields, with probability 1, that $(\beta-\beta_0)^T E[\text{var}(Z\dot{v}(a^T Z)|Y)](\beta-\beta_0) = 0$. Hence by Assumption (E), we have $\beta = \beta_0$, and then $\Lambda(t) = \Lambda_0(t)$ follows. By Theorem 5.8 and Lemma 5.9 in van der Vaart (2002), we conclude that $\hat{\beta}_n \to \beta_0$ and $\|\hat{\Lambda}_n - \Lambda_0\|_2 \to 0$ in probability (almost surely) as $n \to \infty$. By the fact that the density of Y is bounded away from 0, $\|\hat{\Lambda}_n - \Lambda_0\|_2 \to 0$ is equivalent to $\int_\sigma^\tau (\hat{\Lambda}_n(t) - \Lambda_0(t))^2 dt \to 0$ in probability (almost surely). Since $\Lambda_0(\cdot)$ is continuous and strictly monotone, it further implies that $\hat{\Lambda}_n(t) \to \Lambda_0(t)$ in probability (almost surely) for every $t \in (\sigma, \tau)$.

The proof of the rate of convergence follows similarly the proof of Lemma 8.5 in van der Vaart (2002), in which the bracketing number calculation follows the same argument in the proof of Lemma 8.6 in van der Vaart (2002) by using the fact that w is bounded and free of (β, Λ). Details are hence omitted.

Proof of Theorem 2. The proof proceeds similarly as the proof of Theorem 3.4 in Huang (1996) by verifying Conditions A1–A6 in Theorem A.1 with the general relative risk $r(\cdot)$ replacing $\exp(\cdot)$ and $m(\beta, \Lambda, \alpha_0; X) = w(\alpha_0) l(\beta, \Lambda; X)$, where w is bounded and free of (β, Λ), thus is omitted.

Proof of Theorem 3. We show consistency first. Define function $m(\beta, \Lambda, \alpha; X) = w(\alpha)\log\{(p_{\beta,\Lambda} + p_{\beta_0,\Lambda_0})/2\}$. In the proof of Theorem 1 we have already shown that $(\beta_0, \Lambda_0, \alpha_0)$ is the unique maximizer of $Pm(\beta, \Lambda, \alpha_0; X)$. Hence,

$$\sup_{(\beta,\Lambda): d((\beta,\Lambda),(\beta_0,\Lambda_0)) > \delta} Pm(\beta, \Lambda, \alpha_0; X) < Pm(\beta_0, \Lambda_0, \alpha_0; X) \tag{B.1}$$

holds for every $\delta > 0$. By the definition of $(\tilde{\beta}_n, \tilde{\Lambda}_n)$, we have

$$\mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X) \geq \mathbb{P}_n m(\beta_0, \Lambda_0, \hat{\alpha}_n; X) = \mathbb{P}_n m(\beta_0, \Lambda_0, \alpha_0; X) + o_p(1), \tag{B.2}$$

where the equality is obtained by Taylor expansion and the uniform boundedness of $\dot{w}(\alpha)$. By a similar argument as in Lemma 8.9 in van der Vaart (2002), we know that the bracketing numbers of the class of functions $\{m(\beta, \Lambda, \alpha_0; X) : (\beta, \Lambda) \in \Theta\}$ are bounded and hence the class is Glivenko–Cantelli. Thus from (B.1) and (B.2) we have

$$\begin{aligned} 0 &\leq Pm(\beta_0, \Lambda_0, \alpha_0; X) - Pm(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) \\ &= \mathbb{P}_n m(\beta_0, \Lambda_0, \alpha_0; X) - \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) + o_p(1) \\ &\leq \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X) - \mathbb{P}_n m(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) + o_p(1) \\ &= o_p(1), \end{aligned} \tag{B.3}$$

where the last step is again obtained by Taylor expansion and the uniform boundedness of $\dot{w}(\alpha)$. By inequality (B.1), for every $\delta > 0$ we have

$$\left\{d((\tilde{\beta}_n, \tilde{\Lambda}_n), (\beta_0, \Lambda_0)) \geq \delta\right\} \subset \left\{Pm(\tilde{\beta}_n, \tilde{\Lambda}_n, \alpha_0; X) < Pm(\beta_0, \Lambda_0, \alpha_0; X)\right\}$$

with the sequence of the events on the right going to a null event in view of inequalities (B.3), which yields the almost sure (thus in probability) convergence of $(\tilde{\beta}_n, \tilde{\Lambda}_n)$. This argument is taken from the proof of Theorem 5.8 in van der Vaart (2002).

We now show the rate of convergence by applying Theorem 3.2.5 of van der Vaart & Wellner (1996). Let $S_n(\beta, \Lambda) = \mathbb{P}_n w(\hat{\alpha}_n)\ell(\beta, \Lambda; X)$. Clearly $S_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \geq S_n(\beta_0, \Lambda_0)$ by the definition of $(\tilde{\beta}_n, \tilde{\Lambda}_n)$. A Taylor expansion on α at α_0 yields

$$\begin{aligned} S_n(\beta, \Lambda) = \mathbb{P}_n w(\alpha_0)\ell(\beta, \Lambda; X) &+ \mathbb{P}_n \dot{w}^T(\alpha_0)\ell(\beta, \Lambda; X)(\hat{\alpha}_n - \alpha_0) \\ &+ (\hat{\alpha}_n - \alpha_0)^T \mathbb{P}_n \ddot{w}(\alpha_n^*)\ell(\beta, \Lambda; X)(\hat{\alpha}_n - \alpha_0), \end{aligned} \tag{B.4}$$

where α_n^* is a point between α_0 and $\hat{\alpha}_n$. Define $\mathbb{M}_n^0(\beta, \Lambda) = \mathbb{P}_n w(\alpha_0)\ell(\beta, \Lambda; X)$, $\mathbb{M}(\beta, \Lambda) = Pw(\alpha_0)\ell(\beta, \Lambda; X)$, and $\mathbb{M}_n(\beta, \Lambda) = \mathbb{M}_n^0(\beta, \Lambda) + P\dot{w}^T(\alpha_0)\ell(\beta, \Lambda; X)(\hat{\alpha}_n - \alpha_0)$. Then by the uniform boundedness of $\ddot{w}$, it is easy to see that the third term on the right hand side of equality (B.4) is $O_p(n^{-1})$. Thus (B.4) becomes

$$S_n(\beta, \Lambda) = \mathbb{M}_n(\beta, \Lambda) + \frac{1}{\sqrt{n}}\{\mathbb{G}_n \dot{w}^T(\alpha_0)\ell(\beta, \Lambda; X)\}(\hat{\alpha}_n - \alpha_0) + O_p(n^{-1}).$$

In a similar way to the proof of Lemma 7.1 in Huang (1996), we can show that the classes of functions $\{\dot{w}(\alpha_0)^{(j)}\ell(\beta, \Lambda; X) : \beta \in \mathcal{B}, \Lambda \in \Phi\}$, $1 \le j \le J$, are Donsker. Hence

$$\sup_{\beta,\Lambda} |\mathbb{G}_n \dot{w}^{(j)}(\alpha_0)\ell(\beta, \Lambda; X)| = O_p(1), \;\; 1 \le j \le J,$$

and we have $S_n(\beta, \Lambda) = \mathbb{M}_n(\beta, \Lambda) + O_p(n^{-1})$. The inequality $S_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \ge S_n(\beta_0, \Lambda_0)$ then implies that $\mathbb{M}_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \ge \mathbb{M}_n(\beta_0, \Lambda_0) - |O_p(n^{-1})|$, which further implies that $\mathbb{M}_n(\tilde{\beta}_n, \tilde{\Lambda}_n) \ge \mathbb{M}_n(\beta_0, \Lambda_0) - |O_p(r_n^{-2})|$ with $r_n = n^{1/3}$. By a similar argument as in the proof of Theorem 3.3 in Huang (1996), which extends to the weighted likelihood at true α_0 without much difficulty, we obtain

$$\begin{aligned}
&E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left|\sqrt{n}(\mathbb{M}_n^0 - \mathbb{M})(\beta, \Lambda) - \sqrt{n}(\mathbb{M}_n^0 - \mathbb{M})(\beta_0, \Lambda_0)\right| \\
&\le C\delta^{1/2}\left(1 + M\tfrac{\delta^{1/2}}{\delta^2\sqrt{n}}\right).
\end{aligned}$$

Together with the triangle inequality, we then have

$$\begin{aligned}
&E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left|\sqrt{n}(\mathbb{M}_n - \mathbb{M})(\beta, \Lambda) - \sqrt{n}(\mathbb{M}_n - \mathbb{M})(\beta_0, \Lambda_0)\right| \\
&\le E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left|\sqrt{n}(\mathbb{M}_n^0 - \mathbb{M})(\beta, \Lambda) - \sqrt{n}(\mathbb{M}_n^0 - \mathbb{M})(\beta_0, \Lambda_0)\right| \\
&\quad + E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left|\sqrt{n}(\mathbb{M}_n - \mathbb{M}_n^0)(\beta, \Lambda) - \sqrt{n}(\mathbb{M}_n - \mathbb{M}_n^0)(\beta_0, \Lambda_0)\right| \\
&\le C\delta^{1/2}\left(1 + M\tfrac{\delta^{1/2}}{\delta^2\sqrt{n}}\right) \\
&\quad + \sum_{j=1}^{J} \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left|A^{(j)}(\beta, \Lambda) - A^{(j)}(\beta_0, \Lambda_0)|E\sqrt{n}|\hat{\alpha}_{nj} - \alpha_{0j}\right|,
\end{aligned} \tag{B.5}$$

where $A^{(j)}$ is the jth component of $P\dot{w}(\alpha_0)\ell(\cdot, \cdot; X)$. Based on the assumptions on model (3) and the uniform boundedness of $\dot{w}(\alpha_0)$ and $\dot{r}(\cdot)$, we know that for $1 \le j \le J$,

$$\begin{aligned}
|A^{(j)}(\beta, \Lambda) - A^{(j)}(\beta_0, \Lambda_0)| &= |P\dot{w}^{(j)}(\alpha_0)\{\ell(\beta, \Lambda; X) - \ell(\beta_0, \Lambda_0; X)\}| \\
&\le C_j[|\beta - \beta_0| + \{P(\Lambda(Y) - \Lambda_0(Y))^2\}^{1/2}] \\
&= C_j d((\beta, \Lambda), (\beta_0, \Lambda_0)) \\
&\le C_j\delta
\end{aligned}$$

for some constant C_j. Together with the boundedness of $\sup_n E\sqrt{n}|\hat{\alpha}_{nj} - \alpha_{0j}|$, the above inequality implies that the summation term in (B.5) is bounded by $K\delta \le K\delta^{1/2}(1 + M\delta^{1/2}/(\delta^2\sqrt{n}))$ for a constant K and sufficiently small δ. Hence,

$$\begin{aligned}
&E \sup_{d((\beta,\Lambda),(\beta_0,\Lambda_0))<\delta} \left|\sqrt{n}(\mathbb{M}_n - \mathbb{M})(\beta, \Lambda) - \sqrt{n}(\mathbb{M}_n - \mathbb{M})(\beta_0, \Lambda_0)\right| \\
&\le C^*\delta^{1/2}\left(1 + M\tfrac{\delta^{1/2}}{\delta^2\sqrt{n}}\right)
\end{aligned}$$

for a constant C^*.

Finally, the inequality $\mathbb{M}(\beta, \Lambda) - \mathbb{M}(\beta_0, \Lambda_0) \leq -Cd^2((\beta, \Lambda), (\beta, \Lambda)_0)$ can be established as in Lemma 8.8 in van der Vaart (2002). Thus, the conditions of Theorem 3.2.5 of van der Vaart & Wellner (1996) are all satisfied with the same function $\phi_n(\delta)$ as that derived in the proof of Theorem 3.3 in Huang (1996). Hence, $(\tilde{\beta}_n, \tilde{\Lambda}_n)$ converges at the $n^{1/3}$ rate.

Proof of Theorem 4. We prove by checking Conditions A1–A6 in Theorem A.1 with $m(\beta, \Lambda, \alpha; X) = w(\alpha) l(\beta, \Lambda; X)$. Condition A1 holds with $\gamma = 1/3$ by Theorem 3. In order to verify A2, we first need to find an $\mathbf{h}^* \in L_2(P)$ such that $P\dot{m}_{12}(\beta_0, \Lambda_0; X)[h] - P\ddot{m}_{22}(\beta_0, \Lambda_0; X)[\mathbf{h}^*, h] = 0$ for all $h \in H$. Because $E(w|X) = 1$, such a condition reduces to the exact same condition for the full data where $w \equiv 1$, hence holds with the $\mathbf{h}^*$ given in (14), which is the least favourable direction for the full data (see Huang, 1996; Murphy & van der Vaart, 2000; van der Vaart, 2002 for details). Furthermore, A is the information matrix for β for the full data, and its non-singularity is guaranteed by Assumption (E). We thus have verified Condition A2. Condition A3 holds automatically because, by $E(w|X) = 1$, $P\dot{m}_1$ and $P\dot{m}_2$ are equal to the expectations of full data scores for β and Λ, and hence equal to 0 at (β_0, Λ_0).

We now verify Condition A4. The first part of A4 holds automatically since we have $\mathbb{P}_n \dot{m}_1(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X) = 0$. For the second part, we define $\xi_0 = \mathbf{h}^* \circ \Lambda_0^{-1}$ with $\mathbf{h}^*$ given in (14). Using the same argument as that in the proof of Theorem 3.4 in Huang (1996) and taking a Taylor expansion with respect to α at α_0, we obtain

$$\mathbb{P}_n \dot{m}_2(\tilde{\beta}_n, \tilde{\Lambda}_n, \hat{\alpha}_n; X)[\mathbf{h}^*] = J_1 + (\hat{\alpha}_n - \alpha_0)^T J_2 + (\hat{\alpha}_n - \alpha_0)^T J_3(\hat{\alpha}_n - \alpha_0),$$

where

$$\begin{aligned} J_1 &= \mathbb{P}_n \left\{ w(\alpha_0) r(\tilde{\beta}_n^T Z)(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \tilde{\Lambda}_n(Y))(\Delta u(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1 - \Delta)) \right\}, \\ J_2 &= \mathbb{P}_n \left\{ \dot{w}(\alpha_0) r(\tilde{\beta}_n^T Z)(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \tilde{\Lambda}_n(Y))(\Delta u(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1 - \Delta)) \right\}, \end{aligned}$$

and

$$J_3 = \mathbb{P}_n \left\{ \ddot{w}(\alpha_n^*) r(\tilde{\beta}_n^T Z)(\xi_0 \circ \Lambda_0(Y) - \xi_0 \circ \tilde{\Lambda}_n(Y))(\Delta u(Y, Z; \tilde{\beta}_n, \tilde{\Lambda}_n) - (1 - \Delta)) \right\}$$

for some α_n^* lying between α_0 and $\widehat{\alpha}_n$. Following the corresponding calculation of $\mathbb{P}_n \dot{m}_2[\mathbf{h}^*]$ in the proof of Theorem 3.4 in Huang (1996) for the general relative risk $r(\cdot)$ satisfying Assumption (C) and bounded $w(\alpha_0)$ and $\dot{w}(\alpha_0)$ that are free of (β, Λ), we can show that both $J_1 = o_p(n^{-1/2})$ and $J_2 = o_p(n^{-1/2})$. It is easy to see that $J_3 = O_p(1)$ by the boundedness assumptions, hence $(\hat{\alpha}_n - \alpha_0)^T J_3(\hat{\alpha}_n - \alpha_0) = o_p(n^{-1/2})$ because $|\hat{\alpha}_n - \alpha_0| = O_p(n^{-1/2})$. Thus we have verified Condition A4.

To verify A5, it suffices to show that the classes of functions

$$\begin{aligned} \Psi_1(\eta) &= \left\{ w(\alpha)\dot{\ell}_1(\beta, \Lambda; x) - w(\alpha_0)\dot{\ell}_1(\beta_0, \Lambda_0; x) : |\alpha - \alpha_0| + |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \alpha \in R^J, \beta \in \mathcal{B}, \Lambda \in \Phi \right\}, \\ \Psi_2(\eta) &= \left\{ w(\alpha)\dot{\ell}_2(\beta, \Lambda; x)[\mathbf{h}^*] - w(\alpha_0)\dot{\ell}_2(\beta_0, \Lambda_0; x)[\mathbf{h}^*] : |\alpha - \alpha_0| + |\beta - \beta_0| + \|\Lambda - \Lambda_0\|_2 \leq \eta, \alpha \in R^J, \beta \in \mathcal{B}, \Lambda \in \Phi \right\} \end{aligned}$$

are Donsker. This follows in a similar way as that in Lemma 7.1 in Huang (1996) again with the fact that w is bounded and free of (β, Λ).

Finally, A6 is verified by Taylor expansions of functions $P\dot{m}_1(\beta, \Lambda, \alpha; X)$ and $P\dot{m}_2(\beta, \Lambda, \alpha; X)[\mathbf{h}^*]$ at $(\beta_0, \Lambda_0, \alpha_0)$. We also have $\mu = 2$ and $\mu\gamma > 1/2$. Thus, we have completed the proof.

A geometric interpretation of the efficiency gain using estimated weights for the missing data problem is given in the following. Let $\dot{\mathcal{P}}_{\Lambda,\alpha}^{\perp}$ be the orthogonal complement of the tangent space of (Λ, α) in $L_2(P)$. Then the influence function of the regular asymptotic linear estimator $\tilde{\beta}_n$ is in $\dot{\mathcal{P}}_{\Lambda,\alpha}^{\perp}$. Since the score function (or equivalently the influence function) of α, which yields $\hat{\alpha}_n$ for data missing at random, is in $\dot{\mathcal{P}}_{\Lambda,\alpha}$ thus orthogonal to $\dot{\mathcal{P}}_{\Lambda,\alpha}^{\perp}$, we know that $\hat{\alpha}_n$ is asymptotically independent of $\tilde{\beta}_n$, which yields the result given by Pierce (1982). For technical details of this simple interpretation, we refer to Bickel et al. (1993), Robins et al. (1994), and Yu & Nan (2006).

BIBLIOGRAPHY

Bickel, P. J., Klaassen, C. A. J., Ritov, Y., & Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*, Springer, New York.

Billingsley, P. (1995). *Probability and Measure*, 3rd ed., Wiley, New York.

Billingsley, P. (1999). *Convergence of Probability Measures*, 2nd ed., Wiley, New York.

Breslow, N. E. & Wellner, J. A. (2007). Weighted likelihood for semiparametric models and two-phase stratified samples, with application to Cox regression. *Scandinavian Journal of Statistics*, 34, 86–102.

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187–220.

Flynn, N. M., Forthal, D. N., Harro, C. D., Judson, F. N., Mayer, K. H., Para, M. F., & the rgp 120 HIV Vaccine Study Group. (2005). Placebo-controlled phase 3 trial of a recombinant glycoprotein 120 vaccine to prevent HIV-1 infection. *Journal of Infectious Diseases*, 191, 654–665.

Gilbert, P. B., Peterson, M. L., Follmann, D., Hudgens, M. G., Francis, D. P., Gurwith, M., Heyward, W. L., Jobes, D. V., Popovic, V., Self, S. G., Sinangil, F., Burke, D., & Berman, P. W. (2005). Correlation between immunologic responses to a recombinant glycoprotein 120 vaccine and incidence of HIV-1 infection in a phase 3 HIV-1 preventive vaccine trial. *Journal of Infectious Diseases*, 191, 666–677.

Groeneboom, P. & Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*, Birkhäuser, Basel.

Huang, J. (1996). Efficient estimation for the proportional hazards model with interval censored data. *Annals of Statistics*, 24, 540–568.

Kalbfleisch, J. D. & Lawless, J. F. (1988). Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*, 7, 149–160.

Kong, L., Cai, J., & Sen, P. K. (2006). Asymptotic results for fitting semiparametric transformation models to failure time data from case-cohort studies. *Statistica Sinica*, 16, 135–151.

Kulich, M. & Lin, D. Y. (2000). Additive hazards regression for case-cohort studies. *Biometrika*, 87, 73–87.

Li, Z., Gilbert, P., & Nan, B. (2008). Weighted likelihood method for grouped survival data in case-cohort studies with application to HIV vaccine trials. *Biometrics*, 64, 1247–1255.

Lu, W. & Tsiatis, A. A. (2006). Semiparametric transformation models for the case-cohort study. *Biometrika*, 93, 207–214.

Ma, S. & Kosorok, M. R. (2005). Robust semiparametric M-estimation and the weighted bootstrap. *Journal of Multivariate Analysis*, 96, 190–217.

Manski, C. F. & Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrica*, 45, P 1977-L 1988.

Murphy, S. A. & van der Vaart, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association*, 95, 449–465.

Nan, B. (2004). Efficient estimation for case-cohort studies. *Canadian Journal of Statistics*, 32, 403–419.

Nan, B., Emond, M. J., & Wellner, J. A. (2004). Information bounds for Cox regression models with missing data. *Annals of Statistics*, 32, 723–753.

Nan, B., Yu, M., & Kalbfleisch, J. D. (2006). Censored linear regression for case-cohort studies. *Biometrika*, 93, 747–762.

Nan, B., Kalbfleisch, J. D., & Yu, M. (2009). Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Annals of statistics*, 37, 2351–2376.

Neyman, J. (1938). Contribution to the theory of sampling human populations. *Journal of the American Statistical Association*, 33, 101–116.

Pierce, D. A. (1982). The asymptotic effect of substituting estimators for parameters in certain types of statistics. *Annals of Statistics*, 10, 475–478.

Præstgaard, J. & Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Annals of Probability*, 21, 2053–2086.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*, 73, 1–11.

Prentice, R. L. & Self, S. G. (1983). Asymptotic distribution theory for Cox-type regression models with general relative risk form. *Annals of Statistics*, 11, 804–813.

Robertson, T., Wright, F. T., & Dykstra, R. L. (1988). *Order-Restricted Statistical Inference*, Wiley, New York.

Robins, J. M., Rotnitzky, A., & Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846–866.

Self, S. G. & Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics*, 16, 64–81.

Skinner, C. J., Holt, D., & Smith, T. M. F. (Eds.) (1989). *Analysis of Complex Surveys*, John Wiley & Sons, New York.

Thomas, D. C. (1981). General relative-risk models for survival time and matched case-control analysis. *Biometrics*, 37, 673–686.

van der Vaart, A. W. (1998). *Asymptotic Statistics*, Cambridge University Press, Cambridge.

van der Vaart, A. W. (2002). Semiparametric Statistics. In *Lectures on Probability and Statistics, Ecole d'été de Saint-Flour XXIX—1999*, P. Bernard, editor. Springer-Verlag, pp. 330–457.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes*, Springer-Verlag, New York.

Wellner, J. A. & Zhang, Y. (2007). Two likelihood-based semiparametric estimation methods for panel count data with covariates. *Annals of Statistics*, 35, 2106–2142.

Yu, M. & Nan, B. (2006). A revisit of semiparametric regression models with missing data. *Statistica Sinica*, 16, 1193–1212.

Zhang, Y., Hua, L., & Huang, J. (2010). A spline-based semiparametric maximum likelihood estimation method for the Cox model with interval-censored data. *Scandinavian Journal of Statistics*, 37, 338–354.

Received 25 October 2010
Accepted 18 April 2011