University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2007 *Paper* 70

Weighted Likelihood Method for Grouped Survival Data in Case-Cohort Studies with Application to HIV Vaccine Trials

Zhiguo Li* Peter B. Gilbert†
Bin Nan‡

*University of Michigan, zhiguo@umich.edu
†Fred Hutchinson Cancer Research Center & University of Washington, pgilbert@scharp.org
‡University of Michigan, bnan@umich.edu

Weighted Likelihood Method for Grouped Survival Data in Case-Cohort Studies with Application to HIV Vaccine Trials

Zhiguo Li, Peter B. Gilbert, and Bin Nan

Abstract

Grouped failure time data arise often in HIV studies. In a recent preventive HIV vaccine efficacy trial, immune responses generated by the vaccine were measured from a case-cohort sample of vaccine recipients, who were subsequently evaluated for the study endpoint of HIV infection at pre-specified follow-up visits. Gilbert et al. (2005) and Forthal et al. (2007) analyzed the association between the immune responses and HIV incidence with a Cox proportional hazards model, treating the HIV infection diagnosis time as a right censored random variable. The data, however, are of the form of grouped failure time data with case-cohort covariate sampling, and we propose an inverse selection probability weighted likelihood method for fitting the Cox model to these data. The method allows covariates to be time-dependent, and uses multiple imputation to accommodate covariate data that are missing at random. We establish asymptotic properties of the proposed estimators, and present simulation results showing their good finite sample performance. We apply the method to the HIV vaccine trial data, showing that higher antibody levels are associated with a lower hazard of HIV infection.

Weighted Likelihood Method for Grouped Survival Data in Case-Cohort Studies with Application to HIV Vaccine Trials

Zhiguo Li[1], Peter Gilbert[2], and Bin Nan[1,*]

[1] Department of Biostatistics, University of Michigan,
Ann Arbor, MI 48109
[2] Statistical Center for HIV/AIDS Research and Prevention,
Fred Hutchinson Cancer Research Center, Seattle, WA 98109
[*] *email*: bnan@umich.edu

April 12, 2007

Summary. Grouped failure time data arise often in HIV studies. In a recent preventive HIV vaccine efficacy trial, immune responses generated by the vaccine were measured from a case-cohort sample of vaccine recipients, who were subsequently evaluated for the study endpoint of HIV infection at pre-specified follow-up visits. Gilbert et al. (2005) and Forthal et al. (2007) analyzed the association between the immune responses and HIV incidence with a Cox proportional hazards model, treating the HIV infection diagnosis time as a right censored random variable. The data, however, are of the form of grouped failure time data with case-cohort covariate sampling, and we propose an inverse selection probability weighted likelihood method for fitting the Cox model to these data. The method allows covariates to be time-dependent, and uses multiple imputation to accommodate covariate data that are missing at random. We establish asymptotic properties of the proposed estimators, and present simulation results showing their good finite sample performance. We apply the method to the HIV vaccine trial data, showing that higher antibody levels are associated

with a lower hazard of HIV infection.

KEY WORDS: Case-cohort design; HIV vaccine trial; Interval censoring; Proportional hazards model; Random dropout; Weighted likelihood.

1 Introduction

Interval censored data arise when failure times are not exactly observed, but instead the two time points within which each failure happens are observed. The time points may be, for instance, the times of clinic visits. Interval censored failure times are commonly seen in practice, for example patients in clinical trials may be monitored for clinical response at a set of visit times. A special case of interval censored failure times occurs when the visit times are fixed in advance and are the same for all subjects. In this case the failure times are grouped into a discrete set of time intervals. For such a data structure, Kalbfleisch and Prentice (1973) and Prentice and Gloeckler (1978), among others, proposed and developed methods for maximum likelihood estimation of the relative risks and survival function in the proportional hazards model (Cox, 1972; Cox, 1975).

In this article, we consider interval censored survival data with fixed and common visit times arising from cohort studies with case-cohort sampling of certain covariates of interest. The case-cohort design was proposed by Prentice (1986) for large cohort studies (e.g., prevention trials) for which the covariates of interest are expensive to collect. In such a design, the covariate values are collected only for those subjects who experience the failure event during the follow-up period and for a subcohort that is randomly sampled from the study cohort. For right censored data, Self and Prentice (1988) derived the asymptotic theory for a pseudo likelihood estimator of the parameters in a general relative risk model, including the proportional hazards model as a special case.

Gilbert et al. (2005) employed the Self-Prentice method to analyze data from the first randomized placebo-controlled Phase 3 trial of a preventive HIV vaccine (Flynn et al., 2005). Forthal et al. (2007) also analyzed these data, using an alternative psuedo likelihood estimator for the Cox model with case-cohort sampling (Estimator II of Borgan et al., 2000). These analyses addressed the objective to evaluate in vaccine recipients the association be-

tween anti-HIV antibody levels generated by the vaccine and subsequent HIV infection. All volunteers in the trial were immunized with vaccine or placebo at months 0, 1, 6, 12, 18, 24 and 30. Volunteers testing negative for HIV infection at month 0 were enrolled, and HIV infection tests were administered at each immunization visit and at the final follow-up visit at month 36. Serum and plasma samples were obtained from all volunteers at the immunization visits as well as at visits 2 weeks after the immunization visits, scheduled for measuring peak immunologic response values. The assays were performed for all vaccine recipients who became HIV infected and for a stratified random sample of the uninfected vaccine recipients, selected after the trial.

For study participants who acquired HIV infection during the study, the infection time can only be determined to be between the dates of the last negative and first positive HIV tests. In both Gilbert et al.'s (2005) and Forthal et al.'s (2007) Cox model analyses of the case-cohort data, the time to infection was approximated by the midpoint of the dates of the last negative and first positive tests. Covariates included the peak immunologic responses that were measured at 2 weeks after each immunization visit and before the first positive HIV test (if any), and the demographic variables geographic region, race, and baseline behavioral risk score (taking integer values from 0 to 7). The peak immunologic response is time-dependent, but treated as a constant between two adjacent visit times. Approximating interval censoring to right censoring, however, may introduce bias in parameter estimation. It is desirable to develop a more general method that takes the interval censoring nature of the failure times into account.

We propose a weighted likelihood approach to fit a proportional hazards model with grouped survival data and case-cohort covariate sampling. The method maximizes the inverse selection probability weighted log likelihood function (or log partial likelihood function). The weighted likelihood approach has been used in other missing data problems; see Breslow

and Wellner (2007) and references cited therein. In our case, the method leads to consistent and asymptotically normal estimators of the parameters, and the variances of the estimators are consistently estimated by sandwich variance estimators. The numerical calculations can be readily carried out via Newton-Raphson iteration. We apply multiple imputation to handle missing immunological responses in the subcohort. We present the proposed methods and asymptotic results in Section 2 and report a simulation study in Section 3. In Section 4 we apply the proposed method to the vaccine trial example and make concluding remarks in Section 5. We provide proofs of the asymptotic properties in the Appendix.

2 The Weighted Likelihood Method

Let T be the underlying time to the event of interest, and C be the underlying censoring time. Let X be a p-dimensional covariate (process). Assume noninformative censoring and C is independent of T given X. In the HIV vaccine trial study, however, neither T nor C is completely observed. Instead, T is either known to be in one of the m fixed time intervals: $(t_0, t_1], (t_1, t_2], \ldots, (t_{m-1}, t_m)$, where $0 = t_0 < t_1 < \ldots < t_{m-1} < t_m = +\infty$, or right censored at a visit time t_j, $1 \leq j \leq m-1$. In either case, X will be observed up to the last observed visit time. The two cases coincide when $j = m-1$.

Suppose we only observe data in the first R_i intervals for subject i, where $1 \leq R_i \leq m-1$; then the subject either experiences an event in the R_ith interval or is right censored at t_{R_i}. Let $\Delta_{ij} = 1$ if the event for the ith subject falls into the jth interval and $\Delta_{ij} = 0$ otherwise, $1 \leq j \leq R_i$, and denote $\Delta_{i,R_i+1} = 1 - \sum_{j=1}^{R_i} \Delta_{ij}$ and $\Delta_i = (\Delta_{i1}, \ldots, \Delta_{i,R_i+1})'$. In fact $\Delta_{ij} = 0$ for all $j < R_i$, but we keep the vector notation Δ_i for ease of technical derivation. Note that R_i is a random variable and the length of Δ_i varies with R_i. Let the covariate be componentwise constant in each of the m time intervals and denote $X_i = (X_{i1}, \ldots, X_{im})$, where X_{ij} is the p-dimensional covariate vector for the ith subject in the jth interval. Assume

that in a full cohort, we would have n i.i.d. observations (Δ_i, R_i, X_i), $1 \le i \le n$, which is equivalent to observing i.i.d. observations $(\Delta_{i,R_i+1}, R_i, X_i)$, $1 \le i \le n$.

Suppose T follows a Cox regression model, i.e., the hazard function can be written as

$$\lambda(t|X(t)) \;=\; \lambda(t)\exp(X(t)'\beta), \tag{1}$$

where $X(t)$ is the p-dimensional covariate vector at time t and $\beta = (\beta_1, \ldots, \beta_p)'$. Let $\Lambda(t)$ be the baseline cumulative hazard function, and denote $\alpha_k = \Lambda(t_k) - \Lambda(t_{k-1})$ and $\gamma_k = \log \alpha_k, k = 1, 2, \ldots, m$, where α_m and γ_m are equal to $+\infty$. Then the conditional probability of the event for the ith subject falling into the jth interval given X_i is

$$P(\Delta_{ij} = 1|X_i) = e^{-\sum_{k=1}^{j-1} e^{\gamma_k + X_{ik}'\beta}}\left(1 - e^{-e^{\gamma_j + X_{ij}'\beta}}\right), 1 \le j \le m.$$

Here for notational convenience we assume that $\sum_{k=1}^{0} e^{\gamma_k + X_{ik}'\beta} = 0$. Note that the above expression only involves covariates observed up to time t_j for a fixed j.

Clearly the pair of random variables (Δ_i, R_i), or equivalently (Δ_{i,R_i+1}, R_i), is completely determined by (T_i, C_i). In particular, the set $\{\Delta_{i,R_i+1} = 0, R_i = j\}$ is equivalent to observing the event in $(t_{j-1}, t_j]$, which in turn is equivalent to the set $\{T_i \in (t_{j-1}, t_j], C_i \ge t_j\}$; and the set $\{\Delta_{i,R_i+1} = 1, R_i = j\}$ is equivalent to censoring the event at time t_j, which in turn is equivalent to the set $\{T_i \ge t_j, C_i \in (t_{j-1}, t_j]\}$. Let δ_i denote the realized vector values of Δ_i. Then by the conditional independence of T_i and C_i given X_i, the conditional probability mass function of (Δ_i, R_i) given X_i can be written as

$$\begin{aligned}
&P(\Delta_i = \delta_i, R_i = j|X_i) \\
&\quad= \; P\Big\{T_i \in (t_{j-1}, t_j], C_i \ge t_j \Big| X_i\Big\}^{1-\delta_{i,j+1}} P\Big\{T_i \ge t_j, C_i \in (t_{j-1}, t_j] \Big| X_i\Big\}^{\delta_{i,j+1}} \\
&\quad= \; \left\{e^{-\sum_{k=1}^{j-1} e^{\gamma_k + X_{ik}'\beta}}\left(1 - e^{-e^{\gamma_j + X_{ij}'\beta}}\right)\right\}^{1-\delta_{i,j+1}} \left\{e^{-\sum_{k=1}^{j} e^{\gamma_k + X_{ik}'\beta}}\right\}^{\delta_{i,j+1}} f(\delta_i, j|X_i) \\
&\quad= \; \prod_{\ell=1}^{j} \left\{e^{-\sum_{k=1}^{\ell-1} e^{\gamma_k + X_{ik}'\beta}}\left(1 - e^{-e^{\gamma_\ell + X_{i\ell}'\beta}}\right)\right\}^{\delta_{i\ell}} \left\{e^{-\sum_{k=1}^{j} e^{\gamma_k + X_{ik}'\beta}}\right\}^{\delta_{i,j+1}} f(\delta_i, j|X_i)
\end{aligned}$$

$$
\begin{aligned}
&= \prod_{\ell=1}^{j+1} \left(e^{-\sum_{k=1}^{\ell-1} e^{\gamma_k + X_{ik}'\beta}} \right)^{\delta_{i\ell}} \left(1 - e^{-e^{\gamma_j + X_{ij}'\beta}} \right)^{\delta_{ij}} f(\delta_i, j|X_i) \\
&\equiv L(\theta|\Delta_i = \delta_i, R_i = j) f(\delta_i, j|X_i), \qquad 1 \le j \le m-1, \qquad (2)
\end{aligned}
$$

where $f(\delta_i, j|X_i) = \{P(C_i \ge t_j|X_i)\}^{1-\delta_{i,j+1}} \{P(t_j < C_i \le t_{j+1}|X_i)\}^{\delta_{i,j+1}}$ does not contain any information about θ and hence can be dropped when constructing the likelihood function for θ. Note that $L_i(\theta) \equiv L(\theta|\Delta_i, R_i)$ above is more complicated than necessary for numerical evaluation. But its current form will be very helpful in deriving asymptotic properties for the proposed estimator, which will be easily seen in the Appendix. Also note that $L_i(\theta)$ reduces to the likelihood contribution of the ith subject in Prentice and Gloeckler (1978).

In case-cohort studies, the covariates are not observed for all subjects. Here we consider the Bernoulli sampling scheme (Manski and Lerman, 1977) for selecting the subcohort. Each subject is examined for a covariate V_i (which can either be part of X_i or be an ancillary variable(s)) that is measured in all subjects (i.e., at phase one), and is then independently selected at phase two into the subcohort with probability $P(i \in SC|V_i) = \pi(V_i)$, where "$SC$" stands for subcohort and $\pi(\cdot)$ is a known function. The covariate X is assembled only for subjects in the subcohort and for those who experience the failure event during follow-up. The data resulting from this sampling scheme preserve an i.i.d. structure and satisfy the missing at random (MAR) assumption (Little and Rubin, 2002), because the probability that the covariate X is missing depends only on V and $\Delta_{i,R_{i+1}}$, which are always observed.

Kulich and Lin (2004) distinguished between "N-estimation" and "D-estimation" for right censored data in case-cohort sampling designs, where N-estimation uses weights that are independent of failure status while D-estimation uses weights that depend on failure status. The main reason for distinguishing these approaches is that the martingale theory applies for N-estimation, but not for D-estimation. This distinction is irrelevant for our methodology for grouped failure time data because it does not have any difficulty in handling failure status

dependent weights.

For the observed data in a case-cohort study, we propose the following weighted likelihood function for making inferences on θ:

$$L_{w,n}(\theta) = \prod_{i=1}^{n} \Big\{ L_i(\theta) \Big\}^{w_i}, \quad \text{where} \quad w_i = (1 - \Delta_{i,R_i+1}) + \frac{I(i \in SC)}{\pi(V_i)} \Delta_{i,R_i+1}, \quad 1 \le i \le n.$$

Clearly the weight w_i depends on the failure status of subject i. It is easily seen that only subjects with completely observed covariates contribute to the weighted likelihood function, and w_i is the inverse of the probability that subject i is selected from the original cohort to have covariate X_i measured. The logarithm of the weighted likelihood function is

$$\begin{aligned} \ell_{w,n}(\theta) &= \sum_{i=1}^{n} w_i \ell_i(\theta) \\ &= \sum_{i=1}^{n} w_i \left\{ - \sum_{j=1}^{R_i+1} \left(\Delta_{ij} \sum_{k=1}^{j-1} e^{\gamma_k + X_{ik}' \beta} \right) + \Delta_{iR_i} \log \left(1 - e^{-e^{\gamma_{R_i} + X_{iR_i}' \beta}} \right) \right\}. \end{aligned} \tag{3}$$

We call the maximizer of $\ell_{w,n}(\theta)$ the weighted likelihood estimator of θ, denoted by $\hat{\theta}_n$, which can be obtained by solving the following weighted log likelihood estimating equation for θ:

$$\frac{\partial}{\partial \theta} \ell_{w,n}(\theta) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial \theta} \ell_i(\theta) = 0. \tag{4}$$

The Newton-Raphson method can be employed to solve the above estimating equation. Denote $h_{ij} = e^{\gamma_j + X_{ij}' \beta}$, $1 \le i \le n$, $1 \le j \le m-1$. The first order derivatives of the weighted likelihood function are $\partial \ell_{w,n}(\theta)/\partial \theta = \sum_{i=1}^{n} w_i \partial \ell_i(\theta)/\partial \theta$, where

$$\begin{aligned} \frac{\partial \ell_i(\theta)}{\partial \beta} &= - \sum_{j=1}^{R_i+1} \left(\Delta_{ij} \sum_{k=1}^{j-1} h_{ik} X_{ik} \right) + \Delta_{iR_i} \frac{h_{iR_i} e^{-h_{iR_i}}}{1 - e^{-h_{iR_i}}} X_{iR_i}, \\ \frac{\partial \ell_i(\theta)}{\partial \gamma_s} &= - \sum_{j=s+1}^{R_i+1} \{ \Delta_{ij} h_{is} I(R_i \ge s) \} + \Delta_{is} \frac{h_{is} e^{-h_{is}}}{1 - e^{-h_{is}}} I(R_i = s), \quad 1 \le s \le m-1. \end{aligned} \tag{5}$$

Let

$$b_{ij} = \frac{h_{ij} e^{-h_{ij}}}{1 - e^{-h_{ij}}} \left(1 - \frac{h_{ij}}{1 - e^{-h_{ij}}} \right), \quad 1 \le i \le n, \ 1 \le j \le m-1.$$

Then the second order derivatives are $\partial^2 \ell_{w,n}(\theta)/\partial\theta\partial\theta' = \sum_{i=1}^n w_i \partial^2 \ell_i(\theta)/\partial\theta\partial\theta'$, where

$$\begin{aligned}
\frac{\partial^2 \ell_i(\theta)}{\partial\beta\partial\beta'} &= -\sum_{j=1}^{R_i+1}\left(\Delta_{ij}\sum_{k=1}^{j-1} h_{ik}X_{ik}X_{ik}'\right) + \Delta_{iR_i}b_{iR_i}X_{iR_i}X_{iR_i}', \\
\frac{\partial^2 \ell_i(\theta)}{\partial\gamma_s^2} &= -\sum_{j=s+1}^{R_i+1}\{\Delta_{ij}h_{is}I(R_i \geq s)\} + \Delta_{is}b_{is}I(R_i = s), \quad 1 \leq s \leq m-1, \\
\frac{\partial^2 \ell_i(\theta)}{\partial\beta\partial\gamma_s} &= -\sum_{j=s+1}^{R_i+1}\{\Delta_{ij}h_{is}X_{is}I(R_i \geq s)\} + \Delta_{is}b_{is}X_{is}I(R_i = s), \\
\frac{\partial^2 \ell_{w,n}(\theta)}{\partial\gamma_s\partial\gamma_t} &= 0, \quad s \neq t.
\end{aligned}$$

Note that the covariates after the R_ith interval do not contribute to the log likelihood function and its derivatives. Define the matrix of the second derivatives as

$$\begin{aligned}
I_n &= \begin{pmatrix} I_{\gamma\gamma,n} & I_{\gamma\beta,n} \\ I_{\gamma\beta,n}' & I_{\beta\beta,n} \end{pmatrix} \\
&= \begin{pmatrix} -\partial^2\ell_{w,n}(\theta)/\partial\gamma\partial\gamma' & -\partial^2\ell_{w,n}(\theta)/\partial\gamma\partial\beta' \\ -\partial^2\ell_{w,n}(\theta)/\partial\beta\partial\gamma' & -\partial^2\ell_{w,n}(\theta)/\partial\beta\partial\beta' \end{pmatrix}.
\end{aligned}$$

The numerical inversion of I_n is necessary in Newton-Raphson iteration, which may be difficult if there are many intervals (m is large). Following the idea of Prentice and Gloeckler (1978) and Finkelstein (1986), however, the inversion can be simplified by using the following equality

$$I_n^{-1} = \begin{pmatrix} I_{\gamma\gamma,n}^{-1} + AB^{-1}A' & -AB^{-1} \\ -B^{-1}A' & B^{-1} \end{pmatrix},$$

where $A = I_{\gamma\gamma,n}^{-1}I_{\gamma\beta,n}, B = I_{\beta\beta,n} - I_{\gamma\beta,n}'I_{\gamma\gamma,n}^{-1}I_{\gamma\beta,n}$, which only involves inverting the p-dimensional matrix B since $I_{\gamma\gamma,n}$ is diagonal. Then the Newton-Raphson method updates values of $\theta = (\gamma', \beta')'$ iteratively via

$$\begin{pmatrix} \gamma^{(k)} \\ \beta^{(k)} \end{pmatrix} = \begin{pmatrix} \gamma^{(k-1)} \\ \beta^{(k-1)} \end{pmatrix} + \left\{ I_n^{-1}\frac{\partial \ell_{w,n}(\theta)}{\partial\theta} \right\}_{\theta=\theta^{(k-1)}}$$

until the algorithm converges; here the superscript (k) represents values in the kth iteration. Note that when the sample size is small, or some time intervals are narrow, there may be

no observed events in an interval, in which case the Newton-Raphson procedure will fail. A simple remedy is to combine such an interval with its neighbor to make the number of events in the combined interval greater than 0.

The dependency of the sampling probabilities on covariates and outcome makes the case-cohort design a biased sampling design. The inverse selection probability weighted estimating equation (4) corrects the bias, however, because by MAR we have

$$E(w_i|\Delta_i, R_i, X_i, V_i) = (1 - \Delta_{i,R_i+1}) + \Delta_{i,R_i+1}\frac{P(i \in SC|V_i)}{\pi(V_i)} = 1, \tag{6}$$

and hence

$$\begin{aligned} E\left\{w_i\frac{\partial \ell_i(\theta)}{\partial \theta}\right\} &= EE\left\{w_i\frac{\partial \ell_i(\theta)}{\partial \theta}\middle|\, \Delta_i, R_i, X_i, V_i\right\} \\ &= E\left\{\frac{\partial \ell_i(\theta)}{\partial \theta}E(w_i|\Delta_i, R_i, X_i, V_i)\right\} \\ &= E\left\{\frac{\partial \ell_i(\theta)}{\partial \theta}\right\} = 0. \end{aligned} \tag{7}$$

A naive approach to the analysis would simply put $w_i = 1$ for all subjects with covariates completely observed and $w_i = 0$ otherwise. We call the corresponding estimator the naive estimator. Since the equality (6) does not hold for all i, in general the naive estimator will be asymptotically biased, which is verified by the simulation study in Section 3.

For full cohort data, Prentice and Gloeckler (1978) provided an intuitive discussion on the asymptotic properties of the maximum likelihood estimator for grouped survival data. We give a set of mild regularity conditions in the following theorem that formally establishes both consistency and asymptotic normality of the weighted likelihood estimator, which includes the maximum likelihood estimator of Prentice and Gloeckler (1978) as a special case. The proof is deferred to the Appendix.

THEOREM 1: *Suppose the parameter space Θ is compact and the true parameter θ_0 is an interior point of Θ. Assume the following conditions hold:*

(i) The covariate X has bounded support.

(ii) The variance matrix of X_{ij} is positive definite for all $1 \leq j \leq m-1$.

(iii) $\pi(V_i) \geq \delta > 0$ for all i and some $\delta > 0$.

(iv) $P(C_i \geq t_{m-1}|X_i) > 0$ with probability 1.

If the maximizer $\hat{\theta}_n$ of $\ell_{w,n}(\theta)$ does not occur on the boundary of Θ, then as $n \to \infty$, $\hat{\theta}_n$ converges to θ_0 in probability, and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a Gaussian random variable with mean zero and variance $\Sigma(\theta_0) = I^{-1}(\theta_0)D(\theta_0)I^{-1}(\theta_0)$, where $I(\theta) = E_{\theta_0}\{\partial^2 \ell_i(\theta)/\partial\theta\partial\theta'\}$ and $D(\theta) = E_{\theta_0}[\{w_i \partial \ell_i(\theta)/\partial\theta\}\{w_i \partial \ell_i(\theta)/\partial\theta\}']$.

Note that the compactness of Θ and the boundedness of X guarantee that the probability of observing an event in each of the m intervals is strictly bounded between 0 and 1.

The asymptotic variance $\Sigma(\theta_0)$ can be consistently estimated by the sandwich estimator

$$\hat{\Sigma}_n(\hat{\theta}_n) = \hat{I}_n^{-1}(\hat{\theta}_n)\hat{D}_n(\hat{\theta}_n)\hat{I}_n^{-1}(\hat{\theta}_n), \tag{8}$$

where $\hat{I}_n(\theta) = n^{-1}\sum_{i=1}^n w_i\{\partial^2 \ell_i(\theta)/\partial\theta\partial\theta'\}$, $\hat{D}_n(\theta) = n^{-1}\sum_{i=1}^n w_i^2\{\partial \ell_i(\theta)/\partial\theta\}\{\partial \ell_i(\theta)/\partial\theta\}'$.

3 Simulation Study

We conducted simulations to assess the performance of the weighted likelihood estimator by comparing the bias, efficiency and coverage properties to other estimators including the maximum likelihood estimator for full cohort data, the naive estimator for case-cohort data, and the Self-Prentice (1988) pseudo likelihood estimator for case-cohort data. The pseudo likelihood estimation is based on approximating interval censoring by right censoring, whereby event times are defined by the midpoint of the left- and right-censoring intervals.

We consider two covariates (X_1, X_2), where the corresponding coefficients are $(1, -1)'$. Note that the subscript of X here denotes covariate component, not an index for study subject as in Section 2. To match the HIV vaccine trial (Flynn et al., 2005), we set the time

origin as 6.5 months post-entry (the time by which the study subjects are "fully immunized") and use six time intervals ($m = 6$) with fixed visit times at months 12, 18, 24, 30, and 36. The covariate X_1 is set to be discrete and time-independent, which takes values 1 and 2 with equal probability. The covariate $X_2 = (X_{21}, X_{22}, X_{23}, X_{24}, X_{25})'$ is specified as a 5-variate random vector corresponding to the five post-immunization visits at months 6.5, 12.5, 18.5, 24.5, 30.5, where X_{2j} is the covariate value of X_2 in the jth interval. The conditional distribution of X_2 given X_1 is normal, i.e., $X_2|X_1 = k \sim N(\mu_k, \Sigma)$, $k = 1, 2$, with $\mu_1 = (0.1, 0.2, 0.3, 0.4, 0.5)'$, $\mu_2 = (0, 0.1, 0.2, 0.3, 0.4)'$, and

$$\Sigma = \begin{pmatrix} 1 & \rho & \rho^2 & \rho^3 & \rho^4 \\ \rho & 1 & \rho & \rho^2 & \rho^3 \\ \rho^2 & \rho & 1 & \rho & \rho^2 \\ \rho^3 & \rho^2 & \rho & 1 & \rho \\ \rho^4 & \rho^3 & \rho^2 & \rho & 1 \end{pmatrix},$$

where $\rho = 0.7$. With this set-up the covariates X_{2j}, $j = 1, \ldots, 5$, are positively correlated following an AR(1) model, and X_1 and X_2 are also correlated.

We choose the cohort size n as 500 or 3000. When $n = 500$, the probability of selecting censored subjects into the subcohort is 0.25 and the baseline hazard is a constant value 0.02; when $n = 3000$, the selection probability is 0.085 for censored subjects and the baseline hazard is a constant value 0.005. With these settings there are approximately 200 completely observed subjects when $n = 500$, among whom about half are failures, and approximately 400 completely observed subjects when $n = 3000$, among whom about 150 are failures. The latter situation resembles the HIV vaccine trial data that will be analyzed in the next section. The survival times are generated from a piecewise exponential distribution specified by model (1) (with $\lambda_0(t) \equiv c$ specified above). Censoring times are generated from a discrete uniform subdistribution at months $(12, 18, 24, 30)$ combined with a truncation at month 36 to yield about 25 early dropouts (prior month 36), similar to what was observed in the HIV study. One thousand simulation runs are conducted under each simulation setting.

For each simulation run, parameter estimates are obtained by solving equation (4) using the Newton-Raphson method. The initial value of β is set to be zero, and the initial value of γ is obtained from the Kaplan-Meier curve $S^{(0)}(\cdot)$, calculated by pushing the failure time to the right end point of the interval in which an event occurs, via $\gamma_j^{(0)} = \log[\log\{S^{(0)}(t_j)\} - \log\{S^{(0)}(t_{j+1})\}]$, $1 \leq j \leq m-1$. Then the variance estimator is calculated from (8), and the 95% Wald confidence interval for each parameter is obtained based on the asymptotic normality. Bias, coverage percentage, the average of the estimated standard deviations, and the empirical standard deviation are calculated from the 1000 simulation runs. Since the parameter of interest is β, only the bias for estimating γ is reported. The relative efficiency of the weighted likelihood estimator of β versus the maximum likelihood estimator (MLE) computed from the full data is calculated by the ratio of empirical variances.

Due to the expense of measuring the antibody responses in the HIV vaccine trial, the antibody level for vaccine recipients who failed was only measured at the beginning of the first interval (at the month 6.5 visit) and at the visit immediately preceding the failure visit, and for censored vaccine recipients it was only measured at month 6.5 and at a randomly selected visit month after month 6.5. Since the missing elements of X for subject i are missing by design, depending only on $\Delta_{i,R_{i+1}}$, the missing mechanism is MAR (Little and Rubin, 2002). To handle this type of missing data, we propose using multiple imputation to fill in the missing components of X.

Specifically, suppose only X_2 can be missing. For each time interval 2 through 5 (excluding the last interval), we impute the missing values of X_2 by random draws from a linear regression model with the covariate in the first interval as the predictor, which is fitted separately for cases and non-cases. For example, to impute missing covariate values in the second interval for cases, we first fit a linear model $X_{22} = c_0 + c_1 X_{21} + \varepsilon$, where $\varepsilon \sim N(0, \sigma^2)$, using all the cases with complete data for X_{22}. After obtaining estimates $\hat{c} = (\hat{c}_0, \hat{c}_1)'$ and $\hat{\sigma}^2$, we

then take a random draw of σ^{*2} from $\hat{\sigma}^2\chi_{n+1}$, where n is the number of subjects included in the linear regression, and c^* and ε^* are random draws from $N(\hat{c}, \sigma^{*2}(A'A)^{-1})$ and $N(0, \sigma^{*2})$, respectively, where A is the design matrix of the linear regression. Finally, we fill in the missing value X_{22} by $\hat{X}_{22} = c_1^* + c_2^* X_{21} + \varepsilon^*$. We construct 10 complete data sets following this procedure. For each imputed data set, we calculate the weighted likelihood estimator of β and its variance estimate, and then combine the 10 sets of results using the method of Little and Rubin (2002) to obtain the final estimate and its variance estimate. Confidence intervals for β are calculated using the t distribution following Little and Rubin (2002).

In addition to evaluating the different methods with no missing components in X, we evaluate the weighted likelihood method with multiple imputation, by coarsening the simulated X_2 covariates to have missing components in the pattern described above. Tables 1 and 2 summarize the simulation results. From Table 1 we see that the weighted likelihood estimators have reasonably small biases. The standard deviation estimators for $\hat{\beta}$ are accurate, which lead to accurate coverage percentages. The multiple imputation method works well. It is not surprising that the weighted likelihood method for case-cohort data is less efficient than the maximum likelihood estimator for the full cohort data. However, under case-cohort sampling the weighted likelihood method is much more efficient than the naive method that uses simple random sampling. In addition, by ignoring the biased sampling nature of the case-cohort sampled data, the naive estimator is clearly biased. The pseudo likelihood method of Self and Prentice (1988) that uses approximated right censored data is also more biased than the weighted likelihood method for grouped survival data. From Table 2 we see that the bias of $\hat{\gamma}$ is severe for both the naive method and the pseudo likelihood method, whereas it is very small for the weighted likelihood method.

4 Analysis of the HIV Vaccine Trial Data

We now analyze the HIV vaccine trial data using the weighted likelihood method to investigate the association between antibody levels and HIV infection. Subjects randomly assigned to the vaccine group received innoculations at months 0, 1, 6, 12, 18, 24 and 30, and plasma and serum samples were obtained at the "peak antibody response" study visits at months 0.5, 1.5, 6.5, 12.5, 18.5, 24.5 and 30.5. Gilbert et al. (2005) and Forthal et al. (2007) studied many types of antibody measurements in the trial. We investigate the newest antibody measurement described in Forthal et al. (2007), which quantitates the degree to which the serum of a vaccine recipient reduces (relative to control serum) the avidity of the binding of soluble CD4 to the GNE8 strain of HIV. We refer to this antibody variable as the GNE8 CD4 avidity level. We focus on measurements taken at month 6.5, 12.5, 18.5, 24.5, and 30.5 to evaluate the relationship between peak GNE8 CD4 avidity levels and the rate of HIV infection. Because this antibody variable was only obtained from vaccine recipients who tested HIV negative at month 6, and the main scientific goal is to evaluate the association in vaccine recipients after they received the third immunization at month 6.5, the time intervals for analysis are $[6.5, 12)$, $[12.5, 18)$, $[18, 24)$, $[24, 30)$, $[30, 36)$, and $[36, \infty)$, where month 36 is the time of the final study visit.

The GNE8 CD4 avidity level is measured for all infected vaccine recipients and for a stratified random sample of uninfected vaccine recipients. Placebo recipients are not used in the analysis because their GNE8 CD4 avidity levels all equal 0. The stratification variable is defined by five demographic subgroups: white low risk men, nonwhite low risk men, low risk women, white higher risk men, and nonwhite higher risk men, with sampling probabilities 0.05, 0.18, 0.03, 0.20, and 0.45, respectively. Here low (higher) risk subjects are those who had baseline behavioral risk score (defined in Flynn et al., 2005) below or equal to (greater than) 2. The entire cohort size of vaccine recipients at the time origin month 6.5 is 3330, of

whom 131 became HIV infected by month 36. The numbers of uninfected vaccine recipients who were sampled for measuring the GNE8 CD4 avidity level were 113, 69, 66, 25, and 4 for the five strata. Among the 277 sampled uninfected vaccine recipients, 254 were right censored at month 36, and 23 subjects were right censored at an earlier visit time.

In addition to the primary covariate of interest peak GNE8 CD4 avidity level, other covariates included in the Cox model analysis are race (white or nonwhite), sex (male or female) and baseline behavioral risk score. The baseline risk score is categorized into three groups: low (< 2), medium (2 or 3), and high (> 3). The peak antibody level is time-dependent, but is assumed constant between two adjacent vaccine shots. Everyone in the case-cohort data set has the peak antibody level measured in the first time interval (at 6.5 months). For every infected subject, the antibody level is also measured in the time interval in which the infection occurs, while for every uninfected subject, the antibody level is measured in a randomly selected time interval in addition to the first time interval. The antibody levels in all other time intervals are missing by design, and hence are MAR.

To handle the missing covariate data, we use the same approach described in Section 3, wherein 10 complete data sets are created by repeatedly imputing all missing antibody values, the weighted likelihood method is applied to each complete data set, and Little and Rubin's (2002) technique is used to compute the final regression parameter estimates and their associated standard deviations and confidence intervals. During the data exploration we found that the contribution of the antibody level in model (1) is monotone, but not linear, with faster increase at lower antibody levels. By trying out a few power transformations of the antibody level, we found the one fifth power transformation seemed to provide an estimated linear effect. Hence we implemented this transformation in the final analysis.

The results are presented in Table 3. We first investigated interactions between antibody level and the other covariates, and none are statistically significant. On main effects, race

and sex effects are not statistically significant, while baseline risk group is highly significant. Compared to the low risk group, the estimated relative hazard of HIV infection for the medium or high risk groups is approximately tripled, controlling for antibody level, race and gender. The GNE8 CD4 avidity levels are significantly inversely associated with HIV infection rate. Note that on their original scale the antibody levels range from 0 to about 0.75, and their transformed values range from 0 to about 0.95. From Table 3 we see that the estimated log relative hazard of infection for every 0.1 unit increase in the one fifth power of antibody level is -0.156 with 95% confidence interval of $(-0.235, -0.076)$, controlling for race, gender and baseline risk score. Transformed back to the original scale, the strength of association is larger at lower values of the antibody level. For example, an antibody level of 0.25 compared to 0 reduces the hazard of HIV infection by about 69.2%; an antibody level of 0.5 compared to 0.25 reduces the hazard by 16.1%; and the antibody level of 0.75 compared to 0.5 reduces the hazard by 10.8%, controlling for race, gender and baseline risk score.

5 Discussion

The case-cohort sampling considered here is independent Bernoulli sampling that yields random sample sizes. The advantage of this sampling scheme is the resulting i.i.d. structure of the data, which leads to parameter estimators with more manageable asymptotic properties. An alternative approach would be sampling without replacement, wherein the number of sampled subjects is fixed. A different proof of the large sample properties needs to be developed for the non-i.i.d. sampling method. The method of Breslow and Wellner (2007) may apply.

It should also be noted that, although the weighted likelihood estimator provides an intuitively reasonable method that can be easily carried out numerically, it is not the most efficient estimator. Efficient estimation will in general involve the joint distribution of co-

variates and high-dimensional integration, and hence is much more complicated, especially when some covariates are continuous. When covariates are discrete, a simpler derivation is possible, but not pursued here.

Acknowledgements

The authors thank VaxGen, Inc. for providing the HIV vaccine trial data. The work of Gilbert was supported by NIH grant 2 RO1 AI054165-04.

References

Borgan, O., Langholz, B., Samuelsen, S. O., Goldstein, L. and Pogoda, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Analysis* **6,** 39-58.

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

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

Cox, D. R. (1975). Partial likelihood. *Biometrika* **62,** 269-276.

Finkelstein, D. M. (1986). A proportional hazards model for interval-censored failure time data. *Biometrics* **42,** 845-854.

Flynn, N. M., Forthal, D. N., Harro, C. D., Judson, F. N., Mayer, K. H., Para, M. F., and 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.

Forthal, D. N., Gilbert, P. B., Landucci, G. and Phan, T. (2007). Recombinant gp120 vaccine-induced antibodies inhibit clinical strains of HIV-1 in the presence of Fc receptor-bearing effector cells and correlate inversely with HIV infection rate. *Journal of Immunology* (in press).

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. and 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.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association* **47,** 663-685.

Kalbfleisch, J. D. and Prentice, R. L. (1973). Marginal likelihoods based on Cox's regression and life model. *Biometrika* **60**, 267-278.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*. John Wiley, New York.

Kulich, M. and Lin, D. Y. (2004). Improving efficiency of relative-risk estimation in case-cohort studies. *Journal of the American Statistical Association* **99,** 832-844.

Little, R. J. A. and Rubin, D. B. (2000). *Statistical Analysis with Missing Data*. John Wiley, New York.

Manski, C. F. and Lerman, S. R. (1977). The estimation of choice probabilities from choice based samples. *Econometrika* **45,** 1977-1988.

Prentice, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease

prevention trials. *Biometrika* **73,** 1-11.

Prentice, R. L. and Gloeckler, L. A. (1978). Regression analysis of grouped survival data with application to breast cancer data. *Biometrics* **34,** 57-67.

Robins, J. M., Rotnitzky, A. and 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. and Prentice, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Annals of Statistics* **16,** 64-81.

Turnbull, B. W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B* **38,** 290-295.

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

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

Appendix: Proof of Theorem 1

The proof of consistency of $\hat{\theta}_n$ is based on Theorem 5.7 of van der Vaart (1998), which can be reduced to the following Lemma 1 that is more relevant to our problem. In the following we omit the word "outer" from outer probability and outer integral, and refer the detailed arguments to van der Vaart and Wellner (1996), Chapter 1.

LEMMA 1: *For i.i.d. observations* $Z_1, \cdots, Z_n$, *let* $M_n(\theta) = n^{-1} \sum_{i=1}^n m_\theta(Z_i)$ *and* $M(\theta) = Em_\theta(Z)$, *where* $\theta \in \Theta \subset R^d$. *Assume that* Θ *is compact,* $M(\theta)$ *is continuous and has a unique maximizer at* θ_0, *and the measurable function* $\theta \mapsto m_\theta(Z)$ *is continuous for every* Z *and dominated by an integrable function. Then any sequence of estimators* $\hat{\theta}_n$ *satisfying* $M_n(\hat{\theta}_n) \geq M_n(\theta_0) - o_P(1)$ *converges in probability to* θ_0 *as* $n \to \infty$.

PROOF: Since Θ is compact and the function $\theta \mapsto m_\theta(Z)$ is continuous for every Z and dominated by an integrable function, the class of functions $\{m_\theta : \theta \in \Theta\}$ is Glivenko-Cantelli (see example 19.8 in van der Vaart, 1998). Hence we have the uniform convergence of $M_n(\theta)$, i.e., $\sup_{\theta \in \Theta} |M_n(\theta) - M(\theta)| \to 0$ in probability as $n \to \infty$. On the other hand, by the compactness of Θ and the fact that the function $M(\theta)$ has a unique maximizer at θ_0, we have $\sup_{\|\theta - \theta_0\| \geq \varepsilon} M(\theta) < M(\theta_0)$ for every $\varepsilon > 0$. Hence the conditions of Theorem 5.7 of van der Vaart (1998) are satisfied, and it follows that $\hat{\theta}_n \to \theta_0$ in probability.

We now apply Lemma 1 to prove the consistency of $\hat{\theta}_n$ in Theorem 1. By Lemma 1, it suffices to show that the class of functions $\{w\ell(\theta) : \theta \in \Theta\}$ are continuous and bounded by an integrable function, and $\mu(\theta) = E_{\theta_0}\{w\ell(\theta)\}$ is continuous and has a unique maximizer at θ_0, where $\ell(\theta)$ is the log likelihood function for one subject with the subscript i suppressed. From (3) we see that $\ell(\theta)$ is continuous and bounded by a constant since γ, β and X_j are all bounded. In addition, w is bounded by Condition (iii). Thus the function $\theta \mapsto w\ell(\theta)$ is uniformly bounded by an integrable function. Then the continuity of $\mu(\theta)$ follows from the dominated convergence theorem. It remains to show that $\mu(\theta)$ has a unique maximizer at θ_0.

Let $\mu^*(\theta) = \mu(\theta) - \mu(\theta_0)$. Denote the joint density of (Δ, r, X) by p_θ. Then for any $\theta \in \Theta$ we have

$$\mu^*(\theta) = E_{\theta_0}\{w\ell(\theta) - w\ell(\theta_0)\}$$

$$
\begin{aligned}
&= E_{\theta_0}\left\{ w \log \frac{p_\theta(\Delta, r, X)}{p_{\theta_0}(\Delta, r, X)} \right\} \\
&= E_{\theta_0}\left\{ \log \frac{p_\theta(\Delta, r, X)}{p_{\theta_0}(\Delta, r, X)} E_{\theta_0}(w|\Delta, r, X, V) \right\} \\
&= E_{\theta_0}\left\{ \log \frac{p_\theta(\Delta, r, X)}{p_{\theta_0}(\Delta, r, X)} \right\} \qquad \text{by (6)} \\
&\leq \log E_{\theta_0}\left\{ \frac{p_\theta(\Delta, r, X)}{p_{\theta_0}(\Delta, r, X)} \right\} \qquad \text{by the Jensen's inequality} \qquad (9) \\
&= \log 1 = 0.
\end{aligned}
$$

Hence $\mu(\theta)$ is maximized at θ_0. Note that the above calculation shows that $\mu^*(\theta)$ is equivalent to the negative Kullback-Leibler divergence and thus less than or equal to 0. Furthermore, since the equality in (9) holds if and only if $p_{\theta_0}(\Delta, r, X) = p_\theta(\Delta, r, X)$ with probability 1, we have that $\mu(\theta) = \mu(\theta_0)$ if and only if $p_{\theta_0}(\Delta, r, X) = p_\theta(\Delta, r, X)$ with probability 1. Denote $\theta_0 = (\gamma_{1,0}, \cdots, \gamma_{m-1,0}, \beta_0')'$. Then by (2) we have $\gamma_k + X_k'\beta = \gamma_{k,0} + X_k'\beta_0$, or equivalently $X_k'(\beta - \beta_0) = \gamma_{k,0} - \gamma_k$, with probability 1, for all k. Since $\mathrm{Var}(X_k) > 0$, we must have $\beta = \beta_0$ and $\gamma_k = \gamma_{k,0}$ for all k, i.e., $\theta = \theta_0$. Therefore, $\mu(\theta)$ has a unique maximizer at θ_0. Thus the consistency of $\hat{\theta}_n$ follows from Lemma 1.

The proof of asymptotic normality of $\hat{\theta}_n$ in Theorem 1 can be done by applying Theorem 5.23 of van der Vaart (1998), which is listed as Lemma 2 in the following for ease of reference.

LEMMA 2: *Let $Z_1, \ldots, Z_n$ be a random sample from some distribution P. For each θ in an open subset of Euclidean space, let $z \mapsto m_\theta(z)$ be a measurable function such that $\theta \mapsto m_\theta(z)$ is differentiable at θ_0 for $P-$almost every z with derivative $\dot{m}_{\theta_0}(z)$ and such that, for every θ_1 and θ_2 in a neighborhood of θ_0 and a measurable function $\dot{m}$ with $E_P\dot{m}^2 < \infty$,*

$$|m_{\theta_1}(z) - m_{\theta_2}(z)| \leq \dot{m}(z)\|\theta_1 - \theta_2\|.$$

Furthermore, assume that the map $\theta \mapsto E_P m_\theta$ admits a second order Taylor expansion at a point of maximum θ_0 with nonsingular symmetric second derivative matrix V_{θ_0}. If

$\sum_{i=1}^n m_{\hat{\theta}_n}(Z_i) \geq \sup_\theta \sum_{i=1}^n m_\theta(Z_i) - o_p(1)$ *and* $\hat{\theta}_n \to_P \theta_0$, *then*

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{m}_{\theta_0}(Z_i) + o_P(1).$$

PROOF: See van der Vaart (1998), page 54.

We introduce some additional notation before proving the asymptotic normality of $\hat{\theta}_n$. We still suppress the subscript i for subject i because we have i.i.d. observations. For a single observation, let $D = 1$ if the subject either has a failure observed or is right censored at t_{m-1} (the last visit time), and $D = 0$ if the subject is right censored at a time earlier than t_{m-1}. We also extend the length of Δ to m if an event is observed (it is m when the failure time is censored at t_{m-1}) by adding $m - r$ zeros to the remaining intervals after the interval thaqt contains the event. Then the likelihood function for the subject can be decomposed as

$$\begin{aligned} L(\theta) &= \left[\prod_{j=1}^m \left\{ e^{-\sum_{k=1}^{j-1} e^{\gamma_k + X_k'\beta}} \left(1 - e^{-e^{\gamma_j + X_j'\beta}} \right) \right\}^{\Delta_j} \right]^D \left\{ e^{-\sum_{j=1}^r e^{\gamma_j + X_j'\beta}} \right\}^{1-D} \\ &\equiv \{L^{(1)}(\theta)\}^D \{L^{(2)}(\theta)\}^{1-D}. \end{aligned}$$

Likewise, the log likelihood function can be written as

$$\ell(\theta) = D\ell^{(1)}(\theta) + (1 - D)\ell^{(2)}(\theta), \tag{10}$$

where $\ell^{(1)}(\theta) = \log L^{(1)}(\theta)$, and $\ell^{(2)}(\theta) = \log L^{(2)}(\theta)$.

We are now in a position to prove the asymptotic normality of $\hat{\theta}_n$ by checking the conditions of Lemma 2. Identify Z and $m_\theta(Z)$ in the lemma with (Δ, r, X) and $w\ell(\theta)$. Obviously the map $z \mapsto m_\theta(z)$ is measurable and $\theta \mapsto m_\theta(z)$ is differentiable at any θ in Θ for every z. By (5) and the boundedness of w, θ and (Δ, r, X), every element of $\dot{m}_\theta(z) = \partial m_\theta(z)/\partial\theta$

is bounded in both θ and z by a common constant, say, C. By the mean value theorem and the Cauchy-Schwartz inequality we have

$$|m_{\theta_1}(z) - m_{\theta_2}(z)| = |\dot{m}_{\theta^*}(z)'(\theta_1 - \theta_2)| \le \|\dot{m}_{\theta^*}(z)\| \cdot \|\theta_1 - \theta_2\| \le (p + m - 1)C\|\theta_1 - \theta_2\|,$$

where θ^* lies on the line segment between θ_1 and θ_2. Hence we can take $\dot{m}(z)$ in Lemma 2 to be $(m + p - 1)C$ and the condition $E_P\dot{m}^2(Z) < \infty$ is automatically satisfied. Since elements in both $\partial m_\theta(z)/\partial\theta$ and $\partial^2 m_\theta(z)/\partial\theta\partial\theta'$ are bounded by integrable functions, by the dominated convergence theorem we can exchange the second order derivative and the expectation. Hence the map $\theta \mapsto E_P m_\theta$ admits a second-order Taylor expansion. Now we only need to show that V_{θ_0} in Lemma 2 is nonsingular.

By (6) we have $E_P m_\theta = E_P\{w\ell(\theta)\} = E_P\ell(\theta)$. Hence $V_{\theta_0} = E_P\{\partial^2\ell(\theta)/\partial\theta\partial\theta'\} = -I(\theta_0)$. Since

$$I(\theta_0) = -E_P\left\{\frac{\partial^2\ell(\theta)}{\partial\theta\partial\theta'}\right\}_{\theta=\theta_0} = E_P\left\{\frac{\partial\ell(\theta)}{\partial\theta}\left(\frac{\partial\ell(\theta)}{\partial\theta}\right)'\right\}_{\theta=\theta_0},$$

if $I(\theta_0)$ singular, then there must exist a nonzero constant real vector α such that $\alpha' I(\theta_0)\alpha = 0$, which implies by (10) that

$$E_P\left\{\alpha'\frac{\partial\ell(\theta)}{\partial\theta}\right\}^2_{\theta=\theta_0} = E_P\left\{D\left(\alpha'\frac{\partial\ell^{(1)}(\theta)}{\partial\theta}\right)^2 + (1-D)\left(\alpha'\frac{\partial\ell^{(2)}(\theta)}{\partial\theta}\right)^2\right\}_{\theta=\theta_0} = 0.$$

Hence $E_P[D\{\alpha'\partial\ell^{(1)}(\theta)/\partial\theta\}^2]_{\theta=\theta_0} = 0$. Again by (10) we have,

$$\begin{aligned}
\left.\frac{\partial\ell^{(1)}(\theta)}{\partial\gamma_s}\right|_{\theta=\theta_0} &= -\sum_{j=s+1}^{m}\Delta_j h_s^0 + \Delta_s\frac{h_s^0 e^{-h_s^0}}{1-e^{-h_s^0}}, \qquad s = 1, 2, \cdots, m-1, \\
\left.\frac{\partial\ell^{(1)}(\theta)}{\partial\beta}\right|_{\theta=\theta_0} &= \sum_{j=1}^{m}\Delta_j\left(-\sum_{k=1}^{j-1}h_k^0 X_k + \frac{h_j^0 e^{-h_j^0}}{1-e^{-h_j^0}}X_j\right),
\end{aligned}$$

where $h_s^0 = e^{\gamma_s + X_s'\beta_s}|_{\theta=\theta_0}, 1 \le s \le m-1$. Hence we have

$$\begin{aligned}
&E_P\left\{D\left(\alpha'\frac{\partial\ell^{(1)}(\theta)}{\partial\theta}\right)^2\right\}_{\theta=\theta_0} = E_P\left\{\sum_{j=1}^{m}D\Delta_j f_j(X)\right\}^2_{\theta=\theta_0} = E_P\left\{\sum_{j=1}^{m}D\Delta_j f_j^2(X)\right\}_{\theta=\theta_0} \\
&= \sum_{j=1}^{m}E_P\left\{P(\Delta_j = D = 1|X)f_j^2(X)\right\} = 0 \qquad (11)
\end{aligned}$$

for some function f_j. Now by (2), (10) and Assumption (iv), we obtain

$$P(\Delta_j = D = 1|X) = e^{-\sum_{k=1}^{j-1} h_k^0} \left(1 - e^{-h_j^0}\right) P(C \geq t_j|X) > 0, \quad j < m,$$

and

$$P(\Delta_m = D = 1|X) = e^{-\sum_{k=1}^{j-1} h_k^0} \left(1 - e^{-h_j^0}\right) P(C \geq t_{m-1}|X) > 0.$$

Hence (11) holds if and only if $f_j(X) = 0$ with probability 1 for all j. Denoting $\alpha = (c_1, \cdots, c_{m-1}, \bar{\alpha}')'$, then we can write

$$\begin{aligned}
\alpha' \frac{\partial \ell^{(1)}(\theta)}{\partial \theta}\bigg|_{\theta=\theta_0} &= \left\{ \sum_{s=1}^{m-1} c_s \frac{\partial \ell^{(1)}(\theta)}{\partial \gamma_s} + \bar{\alpha}' \frac{\partial \ell^{(1)}(\theta)}{\partial \beta} \right\}_{\theta=\theta_0} \\
&= \sum_{s=1}^{m-1} c_s \left\{ -\sum_{j=s+1}^{m} \Delta_j + \frac{\Delta_s e^{-h_s^0}}{1 - e^{-h_s^0}} \right\} h_s^0 \\
&\qquad + \bar{\alpha}' \sum_{j=1}^{m} \Delta_j \left\{ -\sum_{k=1}^{j-1} h_k^0 X_k + \frac{h_j^0 e^{-h_j^0}}{1 - e^{-h_j^0}} X_j \right\}.
\end{aligned}$$

Therefore the coefficient of Δ_1 is $f_1(X) = (c_1 + \bar{\alpha}' X_1) h_1^0 e^{-h_1^0}/(1 - e^{-h_1^0})$. By setting $f_1(X)$ to be 0, we obtain $c_1 + \bar{\alpha}' X_1 = 0$ with probability 1. Since $\mathrm{Var}(X_1) > 0$, this implies $\bar{\alpha} = 0$ and then it follows that $c_1 = 0$. Now $f_2(X)$ becomes $f_2(X) = c_2 h_2^0 e^{-h_2^0}/(1 - e^{-h_2^0})$, so we have $c_2 = 0$. By continuing this procedure we conclude that $c_3 = \cdots = c_{m-1} = 0$. Therefore, we obtain $\alpha = 0$, which contradicts the assumption of nonzero α. This shows that $I(\theta_0)$ must be nonsingular. Then by Lemma 2 and the consistency of $\hat{\theta}_n$ that we have already shown, we obtain

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = I^{-1}(\theta_0) \frac{1}{\sqrt{n}} \sum_{i=1}^{n} w_i \left. \frac{\partial \ell_i(\theta)}{\partial \theta} \right|_{\theta=\theta_0} + o_P(1),$$

and asymptotic normality is guaranteed by the central limit theorem since $w_i\{\partial \ell_i(\theta)/\partial \theta\}$ is bounded and thus square integrable.

Table 1: Summary statistics of simulations, with true parameter values $\beta_1 = 1$ and $\beta_2 = -1$.

Method	Parameter	Bias	Coverage Percentage	Average SD	Empirical SD	Relative efficiency (from empirical variances)
$n = 500$. Mean sample size of completely observed subjects in the case-cohort sample is 200, in which the mean number of censored subjects selected in the subcohort is 100.						
Weighted	β_1	-0.024	0.931	0.270	0.290	0.607
likelihood	β_2	0.027	0.924	0.115	0.129	0.613
Full data	β_1	-0.009	0.937	0.204	0.226	1
MLE	β_2	0.008	0.932	0.093	0.101	1
Naive	β_1	0.247	0.706	0.195	0.221	–
estimator	β_2	-0.138	0.605	0.087	0.103	–
Pseudo	β_1	0.023	0.910	0.262	0.293	–
likelihood	β_2	-0.104	0.803	0.131	0.146	–
Multiple	β_1	0.035	0.956	0.353	0.294	–
imputation	β_2	-0.014	0.948	0.145	0.148	–
$n = 3000$. Mean sample size of completely observed subjects in the case-cohort sample is 400, in which the mean number of censored subjects selected in the subcohort is 250.						
Weighted	β_1	-0.006	0.941	0.217	0.222	0.519
likelihood	β_2	0.013	0.928	0.099	0.108	0.482
Full data	β_1	-0.004	0.948	0.153	0.160	1
MLE	β_2	0.001	0.948	0.066	0.075	1
Naive	β_1	0.307	0.441	0.146	0.156	–
estimator	β_2	-0.200	0.128	0.061	0.067	–
Pseudo	β_1	0.014	0.896	0.203	0.234	–
likelihood	β_2	-0.090	0.816	0.102	0.118	–
Multiple	β_1	-0.042	0.966	0.247	0.225	–
imputation	β_2	-0.002	0.929	0.120	0.128	–

Table 2: Biases for estimation of the γ_i's in the simulations.

	n=500 ($\gamma_i \equiv -2.41$)					n=3000 ($\gamma_i \equiv -3.95$)				
	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	$\hat{\gamma}_3$	$\hat{\gamma}_4$	$\hat{\gamma}_5$
Weighted likelihood	0.01	-0.01	-0.02	-0.00	-0.02	-0.01	-0.01	-0.00	-0.00	-0.01
Full data MLE	-0.02	-0.01	-0.03	-0.03	-0.02	-0.01	-0.00	-0.01	-0.00	-0.00
Naive estimator	0.57	0.64	0.72	0.85	1.04	1.55	1.63	1.76	1.93	2.11
Pseudo likelihood	0.53	0.29	0.30	0.24	0.31	1.60	1.23	1.28	1.28	1.28

Table 3: Estimated log relative hazards (RHs) of HIV infection in the vaccine trial.

	$(\text{Antibody})^{1/5}$	White	Sex	Medium risk score	High risk score
log(RH)	-1.556	-0.105	-1.411	1.265	1.143
95% CI	(-2.351, -0.757)	(-0.655, 0.445)	(-3.581, 0.759)	(0.741, 1.789)	(0.569, 1.717)
P value	0.001	0.708	0.202	0.000	0.000

White: 1 for white, 0 for nonwhite
Sex: 1 for female, 0 for male
Medium risk group: risk score is equal to 2 or 3
High risk group: risk score is greater than 3