Biometrics 64, 1247–1255
December 2008

DOI: 10.1111/j.1541-0420.2008.00998.x

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, Michigan 48109, U.S.A.
[2]Statistical Center for HIV/AIDS Research and Prevention, Fred Hutchinson
Cancer Research Center, Seattle, Washington 98109, U.S.A.
[*]*email:* bnan@umich.edu

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 prespecified follow-up visits. Gilbert et al. (2005, *Journal of Infectious Diseases* **191**, 666–677) and Forthal et al. (2007, *Journal of Immunology* **178,** 6596–6603) 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 often in HIV studies where times to HIV infection are not exactly observed, but instead the two time points within which the infection happens are observed. The time points may be, for instance, the times of clinic visits. These type of data 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, 1975).

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 pseudolikelihood 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 pseudolikelihood estimator for the Cox model with case–cohort sampling (Estimator II of Borgan et al., 2000). These analyses addressed the objective to evaluate the association between anti-HIV antibody levels generated by the vaccine and subsequent HIV infection in vaccine recipients. Trial participants 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. A small proportion of participants dropped out of the study at earlier times. Serum and plasma samples were obtained from all volunteers at the immunization visits as well as at visits 2 weeks after the immunization, 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. Covariates measured on everyone include demographic variables, geographic region, race, and baseline behavioral risk score (taking integer values from 0 to 7).

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. 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 stratified case–cohort covariate sampling, and apply the method to evaluate the association between the newest antibody measurement described in Forthal et al. (2007) and HIV infection. 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, we consider both true weights and estimated weights, where the true weights are calculated by using the true selection probabilities determined by design and the estimated weights are calculated by using sample fractions within strata. Both methods lead to consistent and asymptotically normal estimators of the parameters, and the variances of the estimators can be consistently estimated. As pointed out by many authors including Breslow and Wellner (2007), the method with estimated weights is more efficient. 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 study and make concluding remarks in Section 5. We provide detailed technical derivations and proofs of the asymptotic properties in the Web-based Supplementary Materials.

2. The Weighted Likelihood Method

Consider the general setting of grouped survival data. 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 $0 = t_0 < t_1 < \cdots < t_{m-1} < t_m = +\infty$, or right censored at a visit time t_j, $1 \le j \le m-1$. In either case, X will be observed up to the last observed visit time. The two cases coincide when $j = m-1$. Here C can be assumed to be discrete with values $t_1, \ldots, t_{m-1}$.

Suppose we only observe data in the first R_i intervals for subject i, where $1 \le R_i \le m-1$; 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, $1 \le j \le R_i$, 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. Following Prentice and Gloeckler (1978), we assume that the covariate is componentwise constant in each of the R_i observed time intervals and denote $X_i = (X_{i1}, \ldots, X_{i,R_i})'$, 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 independent and identically distributed (i.i.d.) observations $(\Delta_i,\ R_i,\ X_i)$, $1 \le i \le n$, which is equivalent to observing i.i.d. observations $(\Delta_{i,R_i+1}, R_i, X_i), 1 \le i \le n$. Clearly the pair of random variables $(\Delta_i,\ R_i)$, 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 $\{T_i \ge t_j,\ C_i \in (t_{j-1},\ t_j]\}$.

Suppose T follows a Cox regression model, that is, the hazard function can be written as

$$\lambda(t \,|\, X(t)) = \lambda(t) \exp(X(t)'\beta), \qquad (1)$$

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

$$P(\Delta_{ij} = 1 \,|\, X_i) = e^{-\sum_{k=1}^{j-1} e^{\gamma_k + X'_{ik}\beta}} \left(1 - e^{-e^{\gamma_j + X'_{ij}\beta}}\right) \\ \times P(C_i \ge t_j \,|\, X_i), \quad 1 \le j \le m.$$

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. The above expression can also be obtained by the first-order approximation of the conditional survival probability given X_i for the Cox model with discrete failure times (see Kalbfleisch and Prentice [2002] for details).

By the conditional independence of T_i and C_i given X_i, the conditional probability mass function of $(\Delta_i,\ R_i)$ given X_i can be written as

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

where $f(\delta_i,\ j \,|\, X_i)$ does not contain any information about $\theta \equiv (\gamma_1, \ldots, \gamma_{m-1},\ \beta')'$ and hence can be dropped when constructing the likelihood function for θ. Detailed derivation is given in Web Appendix A. 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 Web Appendices C and D. Also note that $L_i(\theta)$ reduces to the likelihood contribution of the ith subject in Prentice and Gloeckler (1978).

2.1 *Estimation with True Weights*

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 1), and is then independently selected at phase 2 into the subcohort with probability $P(i \in SC \mid V_i) = \pi(V_i)$, 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 whereas 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:

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

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

$$\begin{aligned}\ell_{w,n}(\theta) &= \sum_{i=1}^{n} w_i \ell_i(\theta) \\ &= \sum_{i=1}^{n} w_i \left\{ - \sum_{j=1}^{R_i+1} \left(\Delta_{ij} \sum_{k=1}^{j-1} e^{\gamma_k + X'_{ik}\beta} \right) \right. \\ &\qquad \left. + \Delta_{iR_i} \log\left(1 - e^{-e^{\gamma_{R_i} + X'_{iR_i}\beta}}\right) \right\}. \qquad (2)\end{aligned}$$

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:

$$\frac{\partial}{\partial\theta} \ell_{w,n}(\theta) = \sum_{i=1}^{n} w_i \frac{\partial}{\partial\theta} \ell_i(\theta) = 0. \qquad (3)$$

The Newton–Raphson method can be employed to solve the above estimating equation. 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

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

where $\gamma = (\gamma_1, \ldots, \gamma_{m-1})'$. 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 because $I_{\gamma\gamma,n}$ is diagonal (see Web Appendix B for explicit forms of the derivatives of the weighted log likelihood). 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 zero and assign the covariate value in the neighbor interval to be the one in the combined interval. We do not encounter such a problem in the HIV data analysis.

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 (3) corrects the bias, however, because by MAR we have

$$\begin{aligned} &E(w_i \mid \Delta_i, R_i, X_i, V_i) \\ &\quad = (1 - \Delta_{i,R_i+1}) + \Delta_{i,R_i+1} \frac{P(i \in SC \mid V_i)}{\pi(V_i)} = 1, \qquad (4) \end{aligned}$$

and hence

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

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. Because the equality (4) 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 with true weights that are usually known for a case–cohort design, which includes the maximum likelihood estimator (MLE) of Prentice and Gloeckler (1978) as a special case. The proof is given in Web Appendix C.

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

Then the maximizer $\hat{\theta}_n$ of $\ell_{w,n}(\theta)$ converges to θ_0 in probability as $n \to \infty$, and $\sqrt{n}(\hat{\theta}_n - \theta_0)$ converges in distribution to a Gaussian random variable with mean zero and variance matrix $\Sigma(\theta_0) = I^{-1}(\theta_0)D(\theta_0)I^{-1}(\theta_0)$, 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. Condition (iv) implies that not all subjects drop out before time t_{m-1}. Otherwise t_{m-2} becomes the last time of visit. The asymptotic variance $\Sigma(\theta_0)$ can be consistently estimated by the sandwich estimator

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

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

2.2 *Estimation with Estimated Weights*

Although the sampling probabilities $\pi(V_i)$ are known, using estimated weights in which $\pi(V_i)$ is replaced by its estimator can improve the efficiency of the weighted likelihood estimator (Robins, Rotnitzky, and Zhao, 1994; Breslow and Wellner, 2007). Suppose that all censored subjects are divided into S strata by the variable $V \in \mathcal{V} \equiv \{\nu_1, \ldots, \nu_S\}$, and in this subsection, we denote the true sampling probabilities by $\pi(\nu_s) = p_{0s}$, $1 \leq s \leq S$. Suppose that there are n_s subjects in stratum s, out of whom n_s^* are selected into the subcohort by the independent Bernoulli sampling. We assume that when $n \to \infty$, $n_s/n \to \alpha_s > 0$, $1 \leq s \leq S$. Instead of using the true sampling probabilities $p_0 = (p_{01}, \ldots, p_{0S})'$ in the weight function w, we now replace each p_{0s} with the sampling fraction $\hat{p}_s = n_s^*/n_s, 1 \leq s \leq S$, and set $\hat{\pi}(V_i) = \hat{p}_s$ if $V_i = \nu_s$, $1 \leq s \leq S$. Now the estimated weight function becomes

$$w_i(\hat{p}) = (1 - \Delta_{i,R_{i+1}}) + \frac{I(i \in SC)}{\hat{\pi}(V_i)}\Delta_{i,R_{i+1}}, \quad 1 \leq i \leq n.$$

Denote the maximizer of $\sum_{i=1}^n w_i(\hat{p})\ell_i(\theta)$ by $\tilde{\theta}_n$. The following theorem establishes the consistency and asymptotic normality of $\tilde{\theta}_n$, but with a different asymptotic variance matrix to that of $\hat{\theta}_n$ given in Theorem 1. A detailed proof is given in Web Appendix D.

THEOREM 2: *Under the same conditions in Theorem* 1, *$\tilde{\theta}_n$ is consistent and $\sqrt{n}(\tilde{\theta}_n - \theta_0)$ converges in distribution to a Gaussian random variable with mean zero and variance*

$$\Sigma(\theta_0) - I^{-1}(\theta_0)B(\theta_0, p_0)G_{22}B'(\theta_0, p_0)I^{-1}(\theta_0)$$

as $n \to \infty$, where

$$B(\theta, p) = E_{\theta_0}[\{\partial\ell_i(\theta)/\partial\theta\}\{\partial w_i(p)/\partial p\}'],$$

$$G_{22} = \text{diag}\{p_{01}(1 - p_{01})/\alpha_1, \ldots, p_{0S}(1 - p_{0S})/\alpha_S\},$$

which can be consistently estimated by

$$\hat{B}(\tilde{\theta}_n, \hat{p}) = \frac{1}{n}\sum_{i=1}^n \{\partial\ell_i(\theta)/\partial\theta\}\{\partial w_i(p)/\partial p\}'\Big|_{\theta=\tilde{\theta}_n, p=\hat{p}},$$

$$\hat{G}_{22} = \text{diag}\{n\hat{p}_1(1 - \hat{p}_1)/n_1, \ldots, n\hat{p}_S(1 - \hat{p}_S)/n_S\}.$$

2.3 *Approaches to Handling Missing Covariate Data*

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 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. Because 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 $m - 1$ (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 noncases. For example, to impute missing covariate values in the second interval for cases, we first fit a linear model $X_{22} = c_0 + c_1X_{21} + \varepsilon$, 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).

The above multiple imputation method for the HIV vaccine case–cohort study assumes that, given the baseline covariate, the covariate distribution in time interval $(t_{j-1}, t_j]$, $j = 2, \ldots, m - 2$, for those who had infection in this interval is the same as that for those who had infection between t_j and t_{m-1}. This may not be true and can only be viewed as an approximation. As t_j gets closer to t_{m-1}, the approximation becomes more precise. Note that such an assumption is not imposed for noncases.

A referee recommended an interesting alternative approach to the multiple imputation. Rather than weight subjects, one can weight occasions within subjects. To be specific, by rewriting the log likelihood for subject i as $\ell_i(\theta) = \sum_{j=1}^{R_i+1} Q_j(X_{ij}, \Delta_{ij}; \theta)$, one can estimate θ by maximizing

$$\sum_{i=1}^{n} \sum_{j=1}^{R_i+1} \frac{\xi_{ij}}{\pi_{ij}} Q_j(X_{ij}, \Delta_{ij}; \theta),$$

where $\xi_{ij} = 1$ if X_{ij} is observed and 0 otherwise, and $\pi_{ij} = P(\xi_{ij} = 1 \mid \Delta_i, V_i)$. This approach does not need the assumption underlying multiple imputation and is easy to implement. However, it does not apply to the particular missing pattern in the HIV vaccine case–cohort study because some of the sampling probabilities π_{ij} in time interval j are actually zero. Simulations under the same settings in the following section show that the method gives large biases if the terms involving zero sampling probabilities are eliminated. Hence we do not consider this alternative approach further in this article. We still present it here because it may work well for other suitable applications.

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 MLE for full cohort data, the naive estimator for case–cohort data, and the Self–Prentice (1988) pseudolikelihood estimator for case–cohort data. The pseudolikelihood 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 regression 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 postimmunization 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, that is, $X_2 \mid X_1 = k \sim N(\mu_k, \Sigma)$, $k = 1, 2$, 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 setup the covariates X_{2j}, $j = 1, \ldots, 5$, are positively correlated following a first-order autoregressive process, and X_1 and X_2 are also correlated.

We choose the cohort size n as 200, 500, or 3000. When $n = 200$, the probability of selecting censored subjects into the subcohort is 0.333 and the baseline hazard is a constant value 0.015; 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 90 completely observed subjects when $n = 200$, among whom about 40 are failures; 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 last 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 (3) with estimated weights 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 endpoint 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 the expressions given in Theorems 1 and 2, 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. Because 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 MLE computed from the full data is calculated by the ratio of empirical variances.

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 in Section 2.3. Tables 1 and 2 summarize the simulation results, where weights are estimated by sampling fractions. 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 pseudolikelihood method of Self and

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

Method	Parameter	Bias	Coverage percentage	Average SE	Empirical SE	Relative efficiency (from empirical variances)
$n = 200$. Mean sample size of completely observed subjects in the case–cohort sample is 90, in which the mean number of censored subjects selected in the subcohort is 50						
Weighted	β_1	−0.007	0.963	0.440	0.435	0.636
likelihood	β_2	0.044	0.942	0.203	0.211	0.720
Full data	β_1	−0.007	0.968	0.093	0.347	1
MLE	β_2	0.014	0.956	0.173	0.179	1
Naive	β_1	0.172	0.923	0.372	0.362	—
estimator	β_2	−0.080	0.907	0.175	0.177	—
Pseudolikelihood	β_1	−0.349	0.813	0.131	0.146	—
	β_2	0.360	0.722	0.262	0.293	—
Multiple	β_1	0.008	0.970	0.481	0.457	—
imputation	β_2	0.074	0.924	0.223	0.230	—
$n = 500$. 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	−0.022	0.942	0.295	0.302	0.580
likelihood	β_2	0.026	0.931	0.133	0.136	0.607
Full data	β_1	−0.026	0.955	0.230	0.230	1
MLE	β_2	0.010	0.954	0.108	0.106	1
Naive	β_1	0.218	0.824	0.233	0.239	—
estimator	β_2	−0.128	0.761	0.108	0.108	—
Pseudolikelihood	β_1	−0.261	0.780	0.131	0.146	—
	β_2	0.249	0.675	0.262	0.293	—
Multiple	β_1	0.030	0.964	0.301	0.287	—
imputation	β_2	0.011	0.959	0.145	0.147	—
$n = 3000$. 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	−0.003	0.945	0.208	0.215	0.561
likelihood	β_2	0.016	0.935	0.096	0.106	0.412
Full data	β_1	−0.018	0.948	0.066	0.161	1
MLE	β_2	−0.002	0.940	0.067	0.068	1
Naive	β_1	0.275	0.562	0.156	0.160	—
estimator	β_2	−0.183	0.229	0.067	0.068	—
Pseudolikelihood	β_1	−0.090	0.863	0.102	0.118	—
	β_2	0.099	0.774	0.203	0.234	—
Multiple	β_1	0.028	0.935	0.215	0.227	—
imputation	β_2	0.019	0.920	0.098	0.110	—

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 pseudolikelihood method, whereas it is very small for the weighted likelihood method.

To better illustrate the efficiency gain of the weighted likelihood estimator with estimated weights compared to the estimator with true weights, we generate an auxiliary variable V that is a coarsening of X. Particularly, $V = 1$ if the average of X_2 over the five intervals is less than 1 and $X_1 = 1$; $V = 2$ if the average of X_2 is less than 1 and $X_1 = 2$; $V = 3$ if the average of X_2 is greater than 1 and $X_1 = 1$; and $V = 4$ if the average of X_2 over the five intervals is greater than 1 and $X_1 = 2$. The subcohort is selected by stratified Bernoulli sampling from the four strata defined by V. When $n = 200$, the subcohort sampling probabilities are 0.4, 0.4, 0.7, and 0.7 for the four strata. When $n = 500$, the sampling probabilities are 0.2, 0.2, 0.7, and 0.7. When $n = 3000$, the sampling probabilities are 0.05, 0.05, 0.25, and 0.25. The probabilities are determined such that the numbers of failures and controls selected into the subcohort are approximately the same as in the previous simulation. Results are given in Table 3, which clearly show the advantage of using estimated weights.

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. 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

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

	Weighted likelihood	Full data MLE	Naive estimator	Pseudolikelihood	Multiple imputation
			$n = 200$, $\gamma_i = -2.41$		
γ_1	−0.13	−0.10	0.45	0.28	0.13
γ_2	−0.07	−0.04	0.56	0.31	0.04
γ_3	−0.02	−0.01	0.68	0.42	0.07
γ_4	−0.04	−0.01	0.77	0.33	0.02
γ_5	−0.06	−0.05	0.85	0.24	−0.03
			$n = 500$, $\gamma_i = -2.12$		
γ_1	0.01	−0.02	0.57	0.53	0.06
γ_2	−0.01	−0.01	0.64	0.29	0.03
γ_3	−0.02	−0.03	0.72	0.30	0.09
γ_4	−0.00	−0.03	0.85	0.24	0.03
γ_5	−0.02	−0.02	1.04	0.31	0.02
			$n = 3000$, true $\gamma_i \equiv -3.51$		
γ_1	−0.01	−0.01	1.55	1.60	0.04
γ_2	−0.01	−0.00	1.63	1.23	0.04
γ_3	−0.00	−0.01	1.76	1.28	0.03
γ_4	−0.00	−0.00	1.93	1.28	0.04
γ_5	−0.01	−0.00	2.11	1.28	0.01

Table 3
Comparing the weighted likelihood methods using true weights and estimated weights

	$\beta_1 = 1$				$\beta_2 = -1$			
	Bias	SE1	SE2	Coverage	Bias	SE1	SE2	Coverage
					$n = 200$			
True weights	0.046	0.195	0.212	0.938	−0.033	0.456	0.446	0.959
Estimated weights	0.037	0.185	0.181	0.917	−0.019	0.397	0.390	0.959
					$n = 500$			
True weights	0.020	0.129	0.121	0.939	0.001	0.288	0.278	0.940
Estimated weights	0.014	0.122	0.117	0.939	0.004	0.255	0.243	0.935
					$n = 3000$			
True weights	0.018	0.095	0.087	0.932	0.013	0.203	0.207	0.948
Estimated weights	0.018	0.085	0.080	0.937	0.007	0.158	0.166	0.955

SE1 = empirical standard error; SE2 = average of estimated standard errors.

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 months 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, 18), [18, 24), [24, 30), [30, 36), and [36, ∞), where month 36 is the time of the final study visit. Because there is only one measurement, if not missing, for the peak GNE8 CD4 avidity level in each time interval for each individual, it is reasonable to assume that this measurement is constant in each time interval.

The GNE8 CD4 avidity level was 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. We only consider men in the analysis because only 4 women were included in the case–cohort sample. The stratification variable is defined by four demographic subgroups: white low-risk men, nonwhite low-risk men, white higher-risk men, and nonwhite higher-risk men, with sampling fractions 0.047, 0.176, 0.208, and 0.450, 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 3370, of whom 131 became HIV infected by month 36. Among uninfected vaccine recipients, 115, 73, 71, and 18 were sampled from the four strata for measuring the GNE8 CD4 avidity level. Among the 277 sampled uninfected vaccine recipients, 254 were right censored at month 36, and 23 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

Table 4

Estimated log-relative hazards (RHs) of HIV infection in the vaccine trial

	$(\text{Antibody})^{1/5}$	White	Medium-risk score	High-risk score
log (RH)	−1.204	−0.191	1.249	1.109
95% CI	(−2.027, −0.342)	(−0.736, 0.354)	(0.728, 1.771)	(0.489, 1.728)
P value	0.009	0.492	<0.001	<0.001

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

Cox model analysis are race (white or nonwhite) 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 to be constant between two adjacent vaccine shots. It is measured at time points described at the beginning of Section 2.3.

To handle the missing covariate data we use the multiple imputation approach described in Section 2.3. 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 that 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 4. We first investigated interactions between antibody level and the other covariates, and none are statistically significant. On main effects, the race effect is not statistically significant, whereas 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 and race. 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 4 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 −0.120 with 95% confidence interval of (−0.203, −0.034), controlling for race 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 59.8%; an antibody level of 0.5 compared to 0.25 reduces the hazard by 12.7%; and the antibody level of 0.75 compared to 0.5 reduces the hazard by 8.5%, controlling for race 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 covariates 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.

We assume constant covariates within each time interval for the HIV vaccine case–cohort study. An ideal model without such assumption would require both (1) a model of how the covariate varies in continuous time; and (2) a model of when a failure event occurred in an interval. For the HIV data, we know from past experiments that the antibody levels tend to decline after they are measured (because they are measured at "peak" immunogenicity time points). However, we do not have "trough" values (i.e., measurements on blood samples taken just before another booster immunization). If we did have the trough values, then perhaps a simple parametric model could be incorporated in the Cox model, but without them it does not seem possible. It is of interest to extend the method to relax the constancy assumption, but because it is complicated and cannot be done for the motivating data set, it is beyond the scope of this article.

6. Supplementary Materials

Web Appendices referenced in Section 2 are available under the Paper Information link at the *Biometrics* website `http://www.biometrics.tibs.org`.

Acknowledgements

The authors thank VaxGen, Inc. for providing the HIV vaccine trial data, and Professor Naisyin Wang, the associate editor, and three referees for their helpful comments. The work of Gilbert was supported by NIH grant 2 RO1 AI054165-04 and the work of Nan was supported in part by NSF grant DMS-07-06700.

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* **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.

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* **178,** 6596–6603.

Gilbert, P. B., et al. (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.

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.* New York: Wiley.

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. (2002). *Statistical Analysis with Missing Data,* 2nd edition. New York: Wiley.

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.

Received July 2007. *Revised November* 2007.
Accepted December 2007.