Published in final edited form as:
Biometrics. 2009 March ; 65(1): 135–142. doi:10.1111/j.1541-0420.2008.01055.x.

Case–Cohort Analysis with Accelerated Failure Time Model

Lan Kong and
Department of Biostatistics, Graduate School of Public Health, University of Pittsburgh, Pittsburgh, Pennsylvania 15261, U.S.A

Jianwen Cai
Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, North Carolina 27599, U.S.A

Lan Kong: lkong@pitt.edu

Summary

In a case–cohort design, covariates are assembled only for a subcohort that is randomly selected from the entire cohort and any additional cases outside the subcohort. This design is appealing for large cohort studies of rare disease, especially when the exposures of interest are expensive to ascertain for all the subjects. We propose statistical methods for analyzing the case–cohort data with a semiparametric accelerated failure time model that interprets the covariates effects as to accelerate or decelerate the time to failure. Asymptotic properties of the proposed estimators are developed. The finite sample properties of case–cohort estimator and its relative efficiency to full cohort estimator are assessed via simulation studies. A real example from a study of cardiovascular disease is provided to illustrate the estimating procedure.

Keywords

Accelerated failure time model; Case-cohort design; Stratified simple random sampling; Survival data

1. Introduction

The case–cohort design (Prentice, 1986) provides a more efficient solution for large cohort studies that involve rare diseases and/or expensive exposures. It is especially useful when multiple outcomes are of interest. Under the classical case–cohort design, expensive covariates are assembled only for a randomly selected sample, subcohort, from the entire cohort at the beginning of the study, and any additional cases or failures outside the subcohort. The case–cohort design is feasible when the outcome can be known and the covariate history is potentially accessible for each cohort member. There have been variations of the basic case–cohort sampling scheme to improve the efficiency of the design (e.g., Borgan et al., 2000; Kulich and Lin, 2000). For instance, the Atherosclerosis Risk in Communities study (ARIC Investigators, 1989) used a stratified sampling of the subcohort to investigate the association between genetic risk factors and development of coronary heart disease (CHD) due to the low incidence rate of CHD and expensive bioassay of blood specimens.

Parametric models for case–cohort studies have been studied in Kalbfleisch and Lawless (1988). Statistical methods for fitting case–cohort data with semiparametric survival models have also been developed for the Cox hazards model (e.g., Prentice, 1986; Self and Prentice, 1988; Wacholder et al., 1989; Lin and Ying, 1993; Barlow, 1994; Chen and Lo, 1999), the additive hazards model (Kulich and Lin, 2000), the proportional odds model (Chen, 2001a) and the semiparametric transformation models (Chen, 2001b; Kong, Cai, and Sen, 2004). These semiparametric models either model the hazard function or survival function of failure time. However, it may be more attractive to model the failure time directly in some applications. The accelerated failure time model naturally fulfils this purpose by linearly relating the natural logarithm of the failure time T to the covariates as

$$\log(T_i)=\beta^{'}Z_i+\varepsilon_i, i=1,\ldots,N \tag{1}$$

where β is an unknown $p \times 1$ vector of regression parameters, β' denotes the transpose of β, Z_i is a p-vector of covariates for the ith individual, ε_i's are independent and identically distributed random errors with an unspecified distribution function F. In addition, model (1) is equivalent to $S_z(t) = 1 - F\{\log(t) - \beta' z\}$, implying that the survival probability $S_z(t)$ for the subjects with a particular covariate vector z at time t is the same as the survival probability for the subjects with $z = 0$ at time $t \exp(-\beta' z)$. Kalbfleisch and Lawless (1988) showed how to perform the case–cohort analysis with parametric accelerated failure time model when the distribution of error term was known. The accelerated failure time model provides an important alternative to the Cox proportional hazards model in that the assumption of proportionality of hazards is not required. It also specifies that the covariate has a multiplicative effect on the failure time rather than the hazard function as in the Cox relative risk models. The role of covariate is to alter the rate at which an individual proceeds along the time axis (Kalbfleisch and Prentice, 2002, p. 44). Note that we only consider the time-invariant covariates in this model, although the time-dependent covariates can be incorporated in the accelerated failure time models (Cox and Oakes, 1984, chapter 5, pp. 64–65; Robins and Tsiatis, 1992;Lin and Ying, 1995).

The least-square based Buckley–James estimator (Buckley and James, 1979) and the rank-based estimators (e.g., Prentice, 1978; Ritov, 1990; Tsiatis, 1990; Wei, Ying, and Lin, 1990; Lai and Ying, 1991; Ying, 1993) have been developed for the accelerated failure time model. Chen and Jewell (2001) considered the accelerated failure time model as a special case of a general class of semiparametric hazards regression models. However, the practical use of this model is rare due to the lack of efficient and reliable computational methods. Recently, Jin et al. (2003) provided and justified rigorously a simple method for implementing the rank estimators through a linear programming approach and estimating the corresponding variance with a resampling technique. It is now important to develop the corresponding methodology for case–cohort data. As model (1) is a natural generalization of parametric log-linear models, our methods for case–cohort data are applicable for the usual linear models as well. In Section 2, we propose the estimating procedures for the regression parameters, and study the asymptotic properties of the proposed estimators. We also conduct simulation studies to investigate the performance of case–cohort estimator under practical sample sizes, as well as its efficiency relative to the full cohort estimator in Section 3. A real case–cohort data set from the ARIC study is used for illustration in Section 4.

2. Methods

We first present the method for simple case–cohort design, where the subcohort is sampled by simple random sampling without replacement. We then extend our approach to the stratified case–cohort design where the subcohort is sampled using a stratified simple random sampling scheme. The asymptotic properties of case–cohort estimators are discussed, whereas the technical details are given in the Appendix.

2.1 Case–Cohort Design with Simple Random Sampling

Let $\{T_i, C_i, Z_i\}$ be N independent replicates of $\{T, C, Z\}$, where C is the potential censoring time and is independent of T conditionally on Z. As a result of censoring, we observe the data (X_i, Δ_i, Z_i), where $X_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leq C_i)$ with I(.) being the indicator function. Define $e_i(\beta) = \log(X_i) - \beta' Z_i$, $N_i(\beta; t) = \Delta_i I\{e_i(\beta) \leq t\}$ and $Y_i(\beta; t) = I\{e_i(\beta) \geq t\}$ for $i = 1, \ldots, N$. In full cohort design, the regression parameters can be estimated from the weighted log-rank estimating function (Ying, 1993; Jin et al., 2003)

$$U_\varphi(\beta)=\sum_{i=1}^{N}\Delta_i\varphi\{\beta;e_i(\beta)\}[Z_i-\overline{Z}\{\beta;e_i(\beta)\}], \quad \text{or}$$
$$U_\varphi(\beta)=\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi(\beta;t)\{Z_i-\overline{Z}(\beta;t)\}dN_i(\beta;t), \tag{2}$$

where φ is a possibly data-dependent weight function, $\bar{Z}(\beta; t) = S^{(1)}(\beta; t)/S^{(0)}(\beta; t)$ with $S^{(0)}(\beta;t)=N^{-1}\sum_{j=1}^{N}Y_j(\beta;t), S^{(1)}(\beta;t)=N^{-1}\sum_{j=1}^{N}Y_j(\beta;t)Z_j$. The choices of $\varphi(\beta; t) = 1$ and $\varphi(\beta; t) = S^{(0)}(\beta; t)$ correspond to the log-rank and Gehan statistics, respectively. The estimating function involves a comparison of the covariate of the observed failure versus the mean of those at risk at that time, this fact reveals intuitively why the case–cohort design may remain efficient by sampling the censored observations.

Let β_0 be the true value of regression parameter vector, and $\hat{\beta}_\varphi$ be a root of estimating function (2). Under certain regularity conditions, it has been shown that the random vector $N^{1/2}(\hat{\beta}_\varphi - \beta_0)$ is asymptotically normal with mean zero and covariance matrix $\sum(\beta_0)=A_\varphi^{-1}(\beta_0)B_\varphi(\beta_0)A_\varphi^{-1}(\beta_0)$ (Tsiatis, 1990;Lai and Ying, 1991;Ying, 1993), where

$$A_\varphi(\beta_0)=\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi(\beta_0;t)\{Z_i-\overline{Z}(\beta_0;t)\}^{\otimes 2}\times\{\dot{\lambda}(t)/\lambda(t)\}dN_i(\beta_0;t),$$
$$B_\varphi(\beta_0)=\lim_{N\to\infty}N^{-1}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi^2(\beta_0;t)\times\{Z_i-\overline{Z}(\beta_0;t)\}^{\otimes 2}dN_i(\beta_0;t),$$

$\lambda(.)$ is the common hazard function of the error terms, $\dot{\lambda}(t) = d\lambda(t)/dt$ and for any column vector a, $a^{\otimes 2} = aa'$.

In the case–cohort studies, the covariate vector Z is not completely available for each individual. Suppose we select a subcohort of size n by simple random sampling without replacement from a cohort study that consists of N independent subjects. Let N_0 and n_0 be the numbers of censored observations in the cohort and subcohort, and N_1 and n_1 be the corresponding numbers of failures. We observe the failure status Δ_i for the entire cohort. However, we only observe complete covariates information for the subcohort members and additional failures outside the subcohort. Let $e(\beta) = \log(X) - \beta' Z$. By virtue of the convergence result in Appendix 1 of Wei et al. (1990), we have that for $\beta \in \mathcal{B}$, a compact set of β_0, $N^{-1}U_\varphi(\beta)$ converges almost surely to

$$u(\beta)=\int_{-\infty}^{\infty}\varphi\{\beta;t\}\left\{d\tau_1(\beta;t)-\frac{\mu_1(\beta;t)}{\mu_0(\beta;t)}d\tau_0(\beta;t)\right\}, \tag{3}$$

as $N \rightarrow \infty$, where $\mu_r(\beta; t) = E[Z^{\otimes r} I\{e(\beta) \geq t\}]$, $\tau_r(\beta; t) = E[Z^{\otimes r} I\{e(\beta) \leq t, \Delta = 1\}]$ $(r = 0, 1)$ and $Z^{\otimes 0} = 1$, $Z^{\otimes 1} = Z$. The quantity $\tau_r(\beta; t)$ $(r = 0,1)$ can be estimated in the same way as in the full cohort analysis because only the failures contribute to this term and the covariate information of failures are complete in the case–cohort studies. By the conditional probability principle, we can express $\mu_r(\beta; t)$ as

$$\gamma E[Z^{\otimes r} I\{e(\beta) \geq t\}|\Delta{=}1]+(1-\gamma)E[Z^{\otimes r} I\{e(\beta) \geq t\}|\Delta{=}0],$$

where $\gamma = \Pr\{\Delta = 1\}$ is the failure rate. Then we can estimate the first term using all the failures and second term using the nonfailures in the subcohort. This strategy has been presented in Chen and Lo (1999) for case–cohort analysis with Cox models. Let $\hat{\gamma}$ be an estimator of γ, estimating the expectations in equation (3) with their empirical counterparts from the case–cohort data yields the estimating function

$$\tilde{u}\,(\beta)=\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi\{\beta;t\}\times\left[Z_i-\frac{\widehat{\gamma}/N_1\sum_{j\in R_1}Y_j\{\beta;t\}Z_j+(1-\widehat{\gamma})/n_0\sum_{j\in\tilde{R}_0}Y_j\{\beta;t\}Z_j}{\widehat{\gamma}/N_1\sum_{j\in R_1}Y_j\{\beta;t\}+(1-\widehat{\gamma})/n_0\sum_{j\in\tilde{R}_0}Y_j\{\beta;t\}}\right]dN_i(\beta;t), \tag{4}$$

where R_1 and $\tilde{R}_0$ denote the index sets of failures in the cohort and nonfailures in the subcohort, respectively. In this article, we assume the cohort is well defined, i.e., N and N_1 are known. With γ estimated by N_1/N and the assumption that n_0/N_0 converges to the same limit as n/N does when n_0, N_0, n, and N go to infinity, we have an estimating function that is equivalent to equation (4) as

$$\tilde{U}_\varphi(\beta)=\sum_{i=1}^{N}\Delta_i\varphi\{\beta;e_i(\beta)\}\left[Z_i-\tilde{Z}\,\{\beta;e_i(\beta)\}\right], \tag{5}$$

where $\tilde{Z}(\beta; t) = \tilde{S}^{(1)}(\beta; t)/\tilde{S}^{(0)}(\beta; t)$ with

$$\tilde{S}^{(0)}(\beta;t)=N^{-1}\sum_{j=1}^{N}h_jY_j(\beta;t),$$
$$\tilde{S}^{(1)}(\beta;t)=N^{-1}\sum_{j=1}^{N}h_jY_j(\beta;t)Z_j,$$

and $h_j = \Delta_j + (1 - \Delta_j)\xi_j/p$, where ξ_j is a subcohort indicator, $p = n/N$ is the selection probability of subcohort and converges to a constant $\alpha \in (0, 1)$ as n and N go to infinity. The case–cohort estimating function $\tilde{U}_\varphi(\beta)$ can be considered as a weighted version of the full cohort estimating function, the weight h_j is the inverse of the probability of being selected into the case–cohort. As in the full cohort analysis, the estimating function $\tilde{U}_\varphi(\beta)$ is generally neither continuous nor monotone in β. Thus, we obtain the estimator β_φ by the linear programming approach as described in Jin et al. (2003). For the Gehan-type weight function $\varphi(\beta; t) = S^{(0)}(\beta; t)$, the estimating function $\tilde{U}_\varphi(\beta)$ reduces to

$$\tilde{U}^{G}_{\varphi}(\beta)=\sum_{i=1}^{N}\Delta_i\tilde{S}^{(0)}\{\beta;e_i(\beta)\}[Z_i-\tilde{Z}\{\beta;e_i(\beta)\}]$$
$$=N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta_i(Z_i-Z_j)I\{e_i(\beta)\le e_j(\beta)\}h_j,$$

which is monotone in each component of β (Fygenson and Ritov, 1994), and is the gradient in β of the convex function

$$L_G(\beta)=N^{-1}\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta_i\{e_i(\beta)-e_j(\beta)\}^{-}h_j$$

with $a^{-} = |a|I(a \le 0)$. The minimization of $L_G(\beta)$ can be carried out by linear programming. The corresponding linear function is $\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta_i u_{ij}$ subject to the linear constraints $u_{ij} \ge 0$ and $u_{ij} \ge -\{e_i(\beta) - e_j(\beta)\}h_j(i, j = 1, \ldots, N)$. Equivalently, we may minimize

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\Delta_i|e_i(\beta)-e_j(\beta)|h_j+|M-\beta'\sum_{k=1}^{N}\sum_{l=1}^{N}\Delta_k(Z_l-Z_k)h_l|,$$

where M is an extremely large number. This type of minimization problem can be solved in the same way as the least absolute deviation regression problem. With Gehan estimator being the initial value, an iterative algorithm proposed in Jin et al. (2003) can be used for the general weight functions. The resulting estimator is consistent although the original estimating equation may contain inconsistent roots.

The case–cohort estimating procedure based on the modification of the full cohort estimating function creates some new technical challenges. In addition to the theoretical justifications used in Ying (1993) for the full cohort estimator, we rely on the asymptotic theory of finite population sampling to address the technical problems arising from the sampling of subcohort without replacement in case–cohort studies. The key step is to approximate the case–cohort estimating function by the full cohort counterpart plus an additional part that is asymptotically uncorrelated to the full cohort part. We show in Appendix A that $N^{-1}\tilde{U}_{\varphi}(\beta)$ converges uniformly to the same nonrandom function as the full cohort counterpart $N^{-1}U_{\varphi}(\beta)$ in a compact neighborhood of β_0. Subsequently, the case–cohort estimator $\tilde{\beta}_{\varphi}$ obtained from equations $\tilde{U}_{\varphi}(\beta) = 0$ is consistent following the same arguments in Ying (1993).

The asymptotic normality of full-cohort estimators of regression coefficients was derived by establishing the asymptotic linearity of estimating function around the shrinking neighborhood of true regression parameter vector β_0. By similar arguments, it can be shown that the case–cohort estimating function is asymptotic linear around the neighborhood of true value β_0. Furthermore, the case–cohort estimating function evaluated at β_0 can be represented by the sum of two independently normally distributed random vectors as shown in Appendix B. The proof is in the same spirit as that of Proposition 1 in Self and Prentice (1988) and equation (3.9) in Samuelson (1997). Then by virtue of asymptotic linearity of estimating function around the shrinking neighborhood of β_0, we have that $\sqrt{N}(\tilde{\beta}_{\varphi}-\beta_0)$ is asymptotically normally distributed with mean zero and covariance matrix $\Sigma(\beta_0) + \Gamma(\beta_0)$, where $\Sigma(\beta_0)$ is the covariance matrix for full cohort estimator, and $\Gamma(\beta_0)$ accounts for the extra variability due to the random sampling of subcohort.

The limiting covariance matrix of the case–cohort estimator is as complicated as that in full cohort analysis. The estimation of the hazard function and its derivative is required. Although the resampling method of Jin et al. (2003) performed well in estimating the covariance matrix, it is not expected to be good for case–cohort data because the censoring rate in case–cohort design is often considerably high. Therefore, we adopt the bootstrap procedure used by Wacholder et al. (1989, p. 119) in the Cox regression analysis of case–cohort data. The bootstrap sample is constructed from the original case–cohort sample such that the size of subcohort, number of cases in the subcohort, and the total number of cases in the bootstrap sample are the same as those in the original sample. Specifically, we sample N_1 cases from the original N_1 cases with replacement and assign the first n_1 cases to be in the subcohort and the rest of the cases to be outside the subcohort, then we sample n_0 noncases from the original n_0 noncases in the subcohort with replacement. Let $\tilde{\beta}^{(i)}$ be the case–cohort estimate obtained from the ith bootstrap sample ($i = 1, \ldots, B$), we may estimate the standard error of $\tilde{\beta}_{\varphi}$ by the square root of sample variance of these B estimates. We assess the performance of the bootstrap method in the simulation study.

2.2 Case–Cohort Design with Stratified Simple Random Sampling

Stratified simple random sampling is usually more efficient than simple random sampling when the stratification variables are correlated with the outcome of interest. Suppose the full cohort consists of K strata of sizes $N_1, N_2, \ldots, N_K$, where $N = N_1 + N_2 + \cdots + N_K$. We select from the kth stratum a random sample of n_k subjects into the subcohort. Then the total subcohort size is $n = n_1 + n_2 + \cdots + n_K$. Let $p_k = n_k/N_k$ be the sampling proportion of the subcohort in the kth stratum and assume that p_k converges to a constant $\alpha_k \in (0, 1)$ as $n_k, N_k \to \infty$. For the kth stratum, we define a function $\tilde{U}^{k}_{\varphi}(\beta)$ in the same way as the unstratified estimating function (5), then the estimating function for the stratified case–cohort design can be written as

$$\tilde{U}^{s}_{\varphi}(\beta)=\sum_{k=1}^{K}\tilde{U}^{k}_{\varphi}(\beta)=\sum_{k=1}^{K}\sum_{i=1}^{N_k}\Delta_{ki}\varphi\{\beta;e_{ki}(\beta)\}[Z_{ki}-\tilde{Z}_k\{\beta;e_{ki}(\beta)\}],$$

where $\tilde{Z}_k(\beta;t)=\tilde{S}^{(1)}_{k}(\beta;t)/\tilde{S}^{(0)}_{k}(\beta;t)$ with

$$\tilde{S}^{(0)}_{k}(\beta;t)=N_k^{-1}\sum_{j=1}^{N_k}h_{kj}Y_{kj}(\beta;t),$$
$$\tilde{S}^{(1)}_{k}(\beta;t)=N_k^{-1}\sum_{j=1}^{N_k}h_{kj}Y_{kj}(\beta;t)Z_{kj},$$

and $h_{kj} = \Delta_{kj} + (1 - \Delta_{kj})\xi_{kj}/p_k$, ξ_{kj} is a subcohort indicator with $\sum_{j=1}^{N_k}\xi_{kj}=n_k$. The same strategies were also used in Kulich and Lin (2000) and Borgan et al. (2000) for stratified case–cohort analysis. The stratified case–cohort estimator $\tilde{\beta}^{s}_{\varphi}$ can be obtained by solving $\tilde{U}^{s}_{\varphi}(\beta)=0$ via the same linear programming approach as described previously for simple case–

cohort design. Because the K strata are independent of each other, we may extend the arguments in Section 2.1 to establish the consistency and the asymptotic normality for the stratified case–cohort estimator. To estimate the variance of $\tilde{\beta}^{s}_{\varphi}$, we first conduct the bootstrap sampling within each stratum as previously described and use the stratified bootstrap data along with the estimating function $\tilde{U}^{s}_{\varphi}(\beta)$ to obtain a bootstrap estimator.

3. Numerical Studies

We carried out simulation studies to assess the performance of the proposed estimating procedure with practical sample sizes and to examine the efficiency of the case–cohort design relative to the full cohort design. First we generated 500 full cohort data sets with sample size of 1500 according to the failure time model $\log T = 3 + \beta Z + \varepsilon$, where the covariate Z, following a Bernoulli distribution with a success probability of 0.1, represented a rare exposure of interest, and the error term ε either followed a standard normal distribution, which resulted in a log-normal distribution for failure time T, or a standard extreme value distribution which resulted in an exponential distribution for T. For a prespecified regression parameter $\beta = -1$ or 0, we generated the censoring time from a uniform distribution on $(0, c)$, where the parameter c was chosen such that the proportion of censoring was expected to be 0.9. In other words, we expected to obtain 10% failure rate. We also created a dichotomous stratification variable V based on two parameters $\eta = \Pr(V = 1 \mid Z = 1)$ and $\nu = \Pr(V = 0 \mid Z = 0)$, where η and ν were the sensitivity and specificity of the surrogate V for the true exposure Z. We chose $\eta = \nu = 0.5$, 0.7, and 0.9, with 0.5 corresponding to the case when V and Z were uncorrelated and the higher number indicates the higher correlation between V and Z. We selected the same size of subcohorts from the two strata by simple random sampling without replacement. We set an overall subcohort sampling proportion as 0.11 such that the number of controls in the resulting case–cohort sample was the same as that of cases (i.e., the average case–cohort size is about 300). It is straightforward to see that each cohort member has the same probability to be selected in the subcohort when $\eta = \nu = 0.5$. In other words, we actually have an unstratified case–cohort design in this case.

We obtained the Gehan-type estimator and the log-rank type estimator for each case–cohort data set via the linear programming method and estimated the standard errors of the estimates by bootstrap method based on 200 bootstrap samples. As shown in Table 1, the proposed estimators of the regression parameters are approximately unbiased for all the cases. Moreover, the means of the estimated standard errors are in good agreement with the empirical standard errors, indicating that the bootstrap variance estimator is fairly good. The empirical 95% confidence intervals also have reasonable coverage rates. The relative efficiencies show that most of the simple unstratified case–cohort estimators reach about 50% of the efficiency of full cohort estimators when only about 20% (300 subjects) of the full cohort subjects were included in the case–cohort estimation. The efficiency of the stratified case–cohort design increases as the correlation between V and Z increases, and can be as high as above 70% when V is a very good surrogate for true exposure Z (i.e., $\eta = \nu = 0.9$). As one reviewer pointed out, the censoring mechanism used may be conservative with respect to case–cohort efficiency. If a fixed follow-up time (e.g., end of study) is used as a common censoring time, the covariate mean in the estimating equation can be better estimated, and thus the case–cohort estimator might do even better.

4. Example

We illustrated the estimating method with a data set from the ARIC study. The ARIC study is a population-based cohort study of cardiovascular diseases, enrolling 15,792 participants aged 45 to 64 years old from 1987 to 1989. A subcohort sample, stratified by gender, age group (≤55 or >55), and average carotid thickness (thin/not thin), was selected for ascertainment of genetic risk factors. In the present analysis, we investigated how a genetic polymorphism of glycoprotein (GP)IIIa, also known as Pl$^{A1/A2}$, was associated with the risk of developing CHD during the first visit to the end of year 1993. The platelet Pl^{A2} allele has been proposed to be a potential factor related to platelet aggregation. We estimated the effect of the allele with accelerated failure time model while adjusting for the covariates such as age, gender (=1 if female), carotid thickness (=1 if thin), race (=1 if African American; =0 otherwise), cholesterol (mg/dl, natural logarithm) and cigarette years of smoking (i.e., average number of cigarettes per day multiplied by number of years smoked/1000). After excluding the patients with missing values on any of these covariates, we had a stratified case–cohort sample of size 944. There were 533 patients in the stratified subcohort sample, and a total of 429 CHD cases (18 cases inside the subcohort and 411 cases outside the subcohort).

Table 2 presented the analysis results using Gehan and log-rank weight functions. Note that the positive regression coefficient in the accelerated failure time model (1) implied a longer survival time and the associated variable had a protective effect. On the other hand, negative coefficients implied earlier development of CHD. For example, the Gehan estimate of the coefficient associated with the allele, −0.12, indicated that the subjects who carried the Pl^{A2} allele developed CHD earlier by a factor of exp(0.12) = 1.13 in time as compared to those who did not carry the allele. However, the effect of the allele was not significant. The ARIC investigators reported the similar results based on the Cox proportional hazards model (Aleksic et al., 2000). They found that the hazard ratio associated with the allele was 1.37 with 95% CI=[0.89, 2.11]. Moreover, older African American males with carotid thickness classified as not thin, with higher cholesterol level and longer years of smoking, were associated with earlier development of CHD. All the adjusted covariates except race had statistically significant effects on the time to develop CHD. The estimates of regression parameters based on the Gehan weight function were similar to those based on the log-rank weight function. This indicated that the model appeared to be adequate because two rank estimators with different weight functions should be close to each other if the assumed model was valid (Wei et al., 1990).

5. Conclusion and Discussion

We developed a rank-based estimating equation approach to fit the failure time data from case–cohort studies with an accelerated failure time model. Our method was also valid for the usual semiparametric linear models. We showed that the proposed estimators were consistent and asymptotically normally distributed. For practical use, we demonstrated that the estimators had nice performance under the finite sample size. The simulation results indicated that the efficiency loss of the case–cohort estimator relative to the full cohort estimator remained acceptable as compared to the sample size reduction. Stratified sampling design further improved efficiency when the stratification variable was a good surrogate of the exposure of interest. As stated in Chen and Lo (1999) for the case of Cox model, an improved case–cohort estimator can be obtained for the accelerated failure time model if the disease prevalence γ is known from census statistics of a disease registry. It is also of interest to examine whether or not the time-dependent weighting scheme used in Barlow (1994) may improve the estimation efficiency. In contrast to the case-cohort estimating function proposed recently by Nan et al. (2006) for the accelerated failure time model, we also included failures outside the subcohort in constructing $\tilde{Z}$ in equation (5) and thus, our estimators may be more efficient. Although some easily measured covariates are available for each cohort member, the covariates information of controls outside the subcohort can not be incorporated in our estimation procedure. An estimating function taking into account all the information on these covariates merits further study.

Acknowledgments

The authors would like to thank the associate editor and the referees for their thoughtful comments and constructive suggestions. The authors would also like to thank the staff and participants in the ARIC study for their important contributions. This research was partially supported by the National Institutes of Health grant R01 HL-57444 and the University of Pittsburgh central research development fund.

References

Aleksic N, Juneja H, Folsom AR, Ahn C, Boerwinkle E, Chambless LE, Wu KK. Platelet Pl(A2) allele and incidence of coronary heart disease: Results from the Atherosclerosis Risk In Communities (ARIC) Study. Circulation 2000;102:1901–1905. [PubMed: 11034936]

ARIC Investigators. The Atherosclerosis Risk in Communities (ARIC) study: Design and objectives. American Journal of Epidemiology 1989;129:687–702. [PubMed: 2646917]

Barlow WE. Robust variance estimation for the case-cohort design. Biometrics 1994;50:1064–1072. [PubMed: 7786988]

Borgan Ø, Goldstein L, Langholz B, Pogoda J, Samuelsen SO. Exposure stratified case-cohort designs. Lifetime Data Analysis 2000;6:39–58. [PubMed: 10763560]

Buckley J, James I. Linear regression with censored data. Biometrika 1979;66:429–436.

Chen HY. Weighted semiparametric likelihood method for fitting a proportional odds regression model to data from the case-cohort design. Journal of the American Statistical Association 2001a; 96:1446–1458.

Chen HY. Fitting semiparametric transformation regression models to data from a modified case-cohort design. Biometrika 2001b;88:255–268.

Chen K, Lo S. Case-cohort and case-control analysis with Cox's model. Biometrika 1999;86:755–764.

Chen YQ, Jewell NP. On a general class of semipara-metric hazards regression models. Biometrika 2001;88:687–702.

Cox, DR.; Oakes, D. Analysis of Survival Data. London: Chapman and Hall; 1984.

Fygenson M, Ritov Y. Monotone estimating equations for censored data. Annals of Statistics 1994;22:732–746.

Hájek J. Limiting distributions in simple random sampling from a finite population. Publications of the Mathematical Institute of the Hungarian Academy of Sciences 1960;5:361–374.

Jin Z, Lin DY, Wei LJ, Ying Z. Rank-based inference for the accelerated failure time model. Biometrika 2003;90:341–353.

Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. Statistics in Medicine 1988;7:147–160.

Kalbfleisch, JD.; Prentice, RL. The Statistical Analysis of Failure Time Data. 2. New York: Wiley; 2002.

Kong L, Cai J, Sen PK. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. Biometrika 2004;91:305–319.

Kulich M, Lin DY. Additive hazards regression for case-cohort studies. Biometrika 2000;87:73–87.

Lai TL, Ying Z. Rank regression methods for left-truncated and right-censored data. Annals of Statistics 1991;19:531–556.

Lin DY, Ying Z. Cox regression with incomplete covariate measurements. Journal of the American Statistical Association 1993;88:1341–1349.

Lin DY, Ying Z. Semiparametric inference for the accelerated life model with time-dependent covariates. Journal of Statistical Planning and Inference 1995;44:47–63.

Nan B, Yu M, Kalbfleisch JD. Censored linear regression for case-cohort studies. Biometrica 2006;93:747–762.

Prentice RL. Linear rank tests with right censored data. Biometrika 1978;65:167–179.

Prentice RL. A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986;73:1–11.

Ritov Y. Estimation in a linear regression model with censored data. Annals of Statistics 1990;18:303–328.

Robins J, Tsiatis AA. Semiparametric estimation of an accelerated failure time model with time dependent covariates. Biometrika 1992;79:311–319.

Samuelson SO. A pseudo-likelihood approach to nested case-control studies. Biometrika 1997;84:379–384.

Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for case-cohort studies. Annals of Statistics 1988;16:64–81.

Tsiatis AA. Estimating regression parameters using linear rank tests for censored data. Annals of Statistics 1990;18:354–372.

Wacholder S, Gail MH, Pee D, Brookmeyer R. Alternative variance and efficiency calculations for the case-cohort design. Biometrika 1989;76:117–123.

Wei LJ, Ying Z, Lin DY. Linear regression analysis of censored survival data based on rank test. Biometrika 1990;77:845–851.

Ying Z. A large sample study of rank estimation for censored regression data. Annals of Statistics 1993;21:76–99.

Appendix

A. Approximation of Estimating Function

We may write

$$N^{-1/2}\tilde{U}_{\varphi}(\beta)=N^{-1/2}U_{\varphi}(\beta)+N^{-1/2}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi(\beta;t)\{\overline{Z}(\beta;t)-\tilde{Z}(\beta;t)\}\,dM_i(\beta;t)+N^{-1/2}\int_{-\infty}^{\infty}\varphi(\beta;t)\{\overline{Z}(\beta;t)-\tilde{Z}(\beta;t)\}d\overline{\Lambda}(\beta;t) \quad (A.1)$$

where $M_i(\beta;t)=N_i(\beta;t)-\int_{-\infty}^{t}Y_i(\beta;u)\lambda(u)\,du$ is a martingale process and $\overline{\Lambda}(\beta;t)=\sum_{i=1}^{N}\Lambda_i(\beta;t)$ with $\Lambda_i(\beta;t)=\int_{-\infty}^{t}Y_i(\beta;u)\lambda(u)\,du$. The first term of (A.1) is the estimating function for full cohort data. It follows from the asymptotic convergence result of finite population sampling and some algebraic manipulation that for r = 0, 1,

$$N^{1/2}\tilde{S}^{(r)}(\beta;t)\approx N^{1/2}S^{(r)}(\beta;t)+N^{1/2}(1-\gamma)\times\left\{n_0^{-1}\sum_{j\in\tilde{R}_0}Y_j(\beta;t)Z_j^{\otimes r}-N_0^{-1}\sum_{j\in R_0}Y_j(\beta;t)Z_j^{\otimes r}\right\}$$

when N_1/N and n_1/n converge to γ as n_1, n, N_1, and N go to infinity, where R_0 denotes the index set of all the censored observations in the cohort, and $\tilde{R}_0$ denotes the corresponding set in the subcohort. Define $S_{R_0}^{(r)}(\beta;t)=N_0^{-1}\sum_{j\in R_0}Y_j(\beta;t)Z_j^{\otimes r}$ and $S_{\tilde{R}_0}^{(r)}(\beta;t)=n_0^{-1}\sum_{j\in\tilde{R}_0}Y_j(\beta;t)Z_j^{\otimes r}$, we have that

$$\begin{aligned} N^{1/2}\{\overline{Z}(\beta;t)- \tilde{Z}(\beta;t)\}&=N^{1/2}\left[\left\{\tilde{S}^{(0)}(\beta;t)-S^{(0)}(\beta;t)\}\overline{Z}(\beta;t)+\{S^{(1)}(\beta;t)-\tilde{S}^{(1)}(\beta;t)\right\}\right]\times\{\tilde{S}^{(0)}(\beta;t)\}^{-1}\\ &=(1-\gamma)N^{1/2}\left[\left\{S^{(0)}_{\tilde{R}_0}(\beta;t)-S^{(0)}_{R_0}(\beta;t)\right\}\overline{Z}(\beta;t)-\left\{S^{(1)}_{\tilde{R}_0}(\beta;t)-S^{(1)}_{R_0}(\beta;t)\right\}\right]\times\{\tilde{S}^{(0)}(\beta;t)\}^{-1}+o_p(1)\\ &=(1-\gamma)N^{1/2}\left[\left\{S^{(0)}_{\tilde{R}_0}(\beta;t)-S^{(0)}_{R_0}(\beta;t)\right\}v(\beta;t)-\left\{S^{(1)}_{\tilde{R}_0}(\beta;t)-S^{(1)}_{R_0}(\beta;t)\right\}\right]\times\{s^{(0)}(\beta;t)\}^{-1}+o_p(1), \end{aligned} \tag{A.2}$$

where $s^{(r)}(\beta;t)$ is the limiting function of $S^{(r)}(\beta;t)$ for $r = 0, 1$ and $v(\beta;t) = s^{(1)}(\beta;t)/s^{(0)}(\beta;t)$. The last equality results from the asymptotic property of finite population sampling, i.e., the average of a certain quantity based on a random sample selected without replacement from a finite population converges in probability to its population counterpart. In addition, equation (A.1) implies that

$$N^{-1}\tilde{U}_{\varphi}(\beta)=N^{-1}U_{\varphi}(\beta)+N^{-1}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi(\beta;t)\{\overline{Z}(\beta;t)-\tilde{Z}(\beta;t)\}dM_i(\beta;t)+N^{-1}\int_{-\infty}^{\infty}\varphi(\beta;t)\{\overline{Z}(\beta;t)-\tilde{Z}(\beta;t)\}d\overline{\Lambda}(\beta;t). \tag{A.3}$$

It follows from the similar conditions and arguments in Lemma 1 of Ying (1993) that $S^{(r)}_{R_0}(\beta;t)$ converges uniformly in t to a nonrandom function within a compact region $\mathcal{B}$ of β_0. Moreover, by the asymptotic result of finite population sampling, $S^{(r)}_{\tilde{R}_0}(\beta;t)$ converges to the same limiting function as their population counterpart. Thus, the last two terms of (A.3) converge to zero uniformly for $\beta \in \mathcal{B}$ in view of equation (A.2). This result implies that the case cohort estimating function $N^{-1}\tilde{U}_{\varphi}(\beta)$ converges uniformly in $\beta \in \mathcal{B}$ to the same nonrandom function as the full cohort counterpart $N^{-1}U_{\varphi}(\beta)$.

B. Normality of N−1/2Ũφ(β0)

It follows from (A.1) that

$$\begin{aligned} N^{-1/2}\tilde{U}_{\varphi}(\beta_0)=&N^{-1/2}U_{\varphi}(\beta_0)\\ &+N^{-1/2}\sum_{i=1}^{N}\int_{-\infty}^{\infty}\varphi(\beta_0;t)\times\{\overline{Z}(\beta_0;t)-\tilde{Z}(\beta_0;t)\}dM_i(\beta_0;t)\\ &+N^{-1}\int_{-\infty}^{\infty}\varphi(\beta_0;t)N^{1/2}\{\overline{Z}(\beta_0;t)-\tilde{Z}(\beta_0;t)\}d\overline{\Lambda}(\beta_0;t). \end{aligned} \tag{B.1}$$

The integral with respect to the martingale in the second term is no longer a martingale because the integrand, involving Δ_i's, is not a predictable process. However, we may apply the Skorokhod strong embedding theorem and the Proposition of Kulich and Lin (2000) to show that the second term converges in probability to zero. Note that the quantity $\bar{Z}(\beta_0;t)$ converges uniformly to $v(\beta_0;t) = s^{(1)}(\beta_0;t)/s^{(0)}(\beta_0;t)$, which is defined in equation (A.2). The proposition of Kulich and Lin (2000) implies that the case–cohort counterpart $\tilde{Z}(\beta_0;t)$ also converges uniformly to the same quantity. Define $B_N(t)=N^{-1/2}\sum_{i=1}^{N}M_i(\beta_0;t)$, we may write the second term in equation (B.1) as

$$\int_{-\infty}^{\infty}\varphi(\beta_0;t)\{\overline{Z}(\beta_0;t)-v(\beta_0;t)\}dB_N(t)-\int_{-\infty}^{\infty}\varphi(\beta_0;t)\{\tilde{Z}\,(\beta_0;t)-v(\beta_0;t)\}dB_N(t).$$

By martingale central limit theorem, $B_N\,(t)$ converges weakly to a tight Gaussian process. It then follows from the same arguments in Kulich and Lin (2000) that $\int_{-\infty}^{\infty}\varphi(\beta_0;t)\overline{Z}(\beta_0;t)dB_N(t)$, $\int_{-\infty}^{\infty}\varphi(\beta_0;t)\,\tilde{Z}\,(\beta_0;t)dB_N(t)$ and $\int_{-\infty}^{\infty}\varphi(\beta_0;t)v(\beta_0;t)dB_N(t)$ converge in probability to the same limit $\int_{-\infty}^{\infty}\varphi(\beta_0;t)v(\beta_0;t)dB(t)$. Thus, the second term in (B.1) converges in probability to zero.

By virtue of (A.2), $N^{1/2}\{\bar{Z}\,(\beta_0;\,t)-\tilde{Z}\,(\beta_0;\,t)\}$ converges weakly to a Gaussian process if $N^{1/2}\{S_{\tilde{R}_0}^{(r)}(\beta_0;t)-S_{R_0}^{(r)}(\beta_0;t)\}\;(r=0,1)$ do. Note that $\{S_{\tilde{R}_0}^{(r)}(\beta_0;t)-S_{R_0}^{(r)}(\beta_0;t)\}\;(r=0,1)$ represents the difference in a certain average between the simple random sample and the corresponding population counterpart. Also $Y_j(\beta_0;t)Z_j^{\otimes r}\;(j=1,\ldots,N)$ are monotone in t and bounded if the covariate vector Z is bounded. It follows from the Proposition of Kulich and Lin (2000) that $N^{1/2}\{S_{\tilde{R}_0}^{(r)}(\beta_0;t)-S_{R_0}^{(r)}(\beta_0;t)\}$ converges weakly to a tight Gaussian process. Combining the facts that $N^{-1}\,\bar{\Lambda}(.)$ is a bounded monotone absolutely continuous function and the linear functional of the Gaussian process is Gaussian, the third term of (B.1) follows a normal distribution. The first term of (B.1), $N^{-1/2}U_{\,\varphi}(\beta_0)$, corresponds to the full cohort counterpart and is shown to be asymptotical normal. By the characteristic function approach and extension of Hájek's (1960) central limit theorem, one can show that the first and third terms in (B.1) are mutually independent of each other and jointly converge to a normally distributed random vector. This result also follows from the Proposition 1 of Self and Prentice (1988). Therefore, $N^{-1/2}\tilde{U}_{\varphi}(\beta_0)$ is asymptotically normal.

Table 1

Simulation summary statistics for estimation of regression parameters based on 500 replications and 200 bootstrap samples

				Mean		
Design	**Weight**	**Mean($\tilde{\beta}$)**	**SD($\tilde{\beta}$)**	**SE**	**CP (%)**	**RE**
(a) $\beta = -1$, error term follows $N(0, 1)$ distribution						
SRS	Gehan	−1.01	0.19	0.19	94.0	54.6
	Log rank	−1.00	0.20	0.21	93.2	47.6
Strat1	Gehan	−0.98	0.18	0.18	94.3	63.5
	Log rank	−0.98	0.19	0.19	94.7	55.5
Strat2	Gehan	−0.99	0.16	0.15	93.2	74.0
	Log rank	−0.99	0.16	0.15	94.8	67.8
(b) $\beta = 0$, error term follows $N\ (0, 1)$ distribution						
SRS	Gehan	−0.01	0.24	0.23	94.8	48.2
	Log rank	−0.01	0.26	0.23	93.6	42.1
Strat1	Gehan	0.00	0.21	0.22	95.8	58.8
	Log rank	0.01	0.22	0.22	94.2	51.4
Strat2	Gehan	0.01	0.17	0.16	94.4	79.6
	Log rank	0.01	0.18	0.16	93.8	71.9
(c) $\beta = -1$, error term follows extreme value distribution						
SRS	Gehan	−1.03	0.36	0.37	94.0	53.8
	Log rank	−1.03	0.35	0.34	93.2	48.2
Strat1	Gehan	−1.03	0.34	0.35	94.2	59.7
	Log rank	−1.03	0.32	0.32	93.8	55.1
Strat2	Gehan	−0.99	0.29	0.28	93.4	82.1
	Log rank	−0.99	0.27	0.25	93.0	72.4
(d) $\beta = 0$, error term follows extreme value distribution						
SRS	Gehan	0.02	0.41	0.43	96.4	58.3
	Log rank	0.02	0.40	0.40	95.4	54.3
Strat1	Gehan	−0.00	0.40	0.40	95.4	63.4
	Log rank	0.00	0.40	0.37	94.0	59.7
Strat2	Gehan	0.00	0.34	0.31	94.2	77.7

				Mean		
Design	**Weight**	**Mean($\tilde{\beta}$)**	**SD($\tilde{\beta}$)**	**SE**	**CP (%)**	**RE**
	Log rank	0.01	0.33	0.28	93.8	75.9

SRS, Simple random sampling case–cohort design; Strat1, stratified case–cohort design with $\eta = \nu = 0.7$; Strat2, stratified case–cohort design with $\eta = \nu = 0.9$; SD, sample standard deviation; SE, estimated standard error; CP, empirical coverage probability of the 95% confidence intervals; RE, empirical relative efficiencies of case–cohort estimators, calculated by the ratio of sample variances of parameter estimates with the full cohort design being a reference.

Table 2

Case–cohort analysis of ARIC data with accelerated failure time model

Covariates	Estimate	SE	95%CI
(a) Gehan weight function			
Pl^{A2} Allele (=1 if present)	−0.12	0.21	(−0.54, 0.29)
Age	−0.07	0.01	(−0.09, −0.05)
Race (=1 if African American)	−0.28	0.21	(−0.69, 0.13)
Log(cholesterol)	−1.87	0.52	(−2.89, −0.86)
Years of smoking	−1.45	0.31	(−2.05, −0.85)
Gender(=1 if female)	1.06	0.11	(0.85, 1.28)
Carotid(=1 if thin)	1.05	0.09	(0.87, 1.22)
(b) Log-rank weight function			
Pl^{A2} Allele	−0.10	0.25	(−0.59, 0.40)
Age	−0.06	0.01	(−0.09, −0.04)
Race	−0.24	0.27	(−0.76, 0.29)
Log(cholesterol)	−1.88	0.68	(−3.20, −0.55)
Years of smoking	−1.24	0.41	(−2.05, −0.44)
Gender	1.02	0.14	(0.75, 1.30)
Carotid	1.02	0.09	(0.85, 1.19)

SE, estimated standard error; CI, confidence interval.