Biostatistics (2006), **7**, 3, *pp.* 486–502
doi:10.1093/biostatistics/kxj021
Advance Access publication on February 24, 2006

Efficient semiparametric estimation of haplotype-disease associations in case–cohort and nested case–control studies

D. ZENG, D. Y. LIN*

Department of Biostatistics, CB# 7420, University of North Carolina, Chapel Hill, NC 27599-7420, USA
lin@bios.unc.edu

C. L. AVERY, K. E. NORTH

Department of Epidemiology, CB# 7435, University of North Carolina, Chapel Hill, NC 27599-7420, USA

M. S. BRAY

Department of Pediatrics, Baylor College of Medicine, Houston, TX 77030, USA

SUMMARY

Estimating the effects of haplotypes on the age of onset of a disease is an important step toward the discovery of genes that influence complex human diseases. A haplotype is a specific sequence of nucleotides on the same chromosome of an individual and can only be measured indirectly through the genotype. We consider cohort studies which collect genotype data on a subset of cohort members through case–cohort or nested case–control sampling. We formulate the effects of haplotypes and possibly time-varying environmental variables on the age of onset through a broad class of semiparametric regression models. We construct appropriate nonparametric likelihoods, which involve both finite- and infinite-dimensional parameters. The corresponding nonparametric maximum likelihood estimators are shown to be consistent, asymptotically normal, and asymptotically efficient. Consistent variance–covariance estimators are provided, and efficient and reliable numerical algorithms are developed. Simulation studies demonstrate that the asymptotic approximations are accurate in practical settings and that case–cohort and nested case–control designs are highly cost-effective. An application to a major cardiovascular study is provided.

Keywords: Age of onset; Association studies; Censoring; Haplotype effects; Nonparametric likelihood; Proportional hazards; Semiparametric efficiency; Single nucleotide polymorphisms; Survival data.

*To whom correspondence should be addressed.

1. Introduction

Complex human diseases, such as cancer, diabetes, schizophrenia, and coronary heart disease (CHD), are affected by multiple genetic and environmental factors. Recent sequencing of the human genome and advances in genotyping technologies have spurred an enormous interest in genetic association studies which explore the relationships between complex diseases and single nucleotide polymorphisms (SNPs). SNPs are single-base variations in the genetic code that occur about every 1000 bases along the 3 billion bases of the human genome. A specific combination of nucleotides at a series of nearby SNPs on the same chromosome of an individual is called a haplotype. The use of haplotypes can yield more powerful tests of genetic associations than the use of single SNPs, especially when the disease-predisposing SNPs are not directly measured or when there are strong interactions of multiple SNPs on the same chromosome (Akey *et al.*, 2001; Morris and Kaplan, 2002; Schaid *et al.*, 2002; Zaykin *et al.*, 2002; Schaid, 2004).

Current genotyping technologies cannot separate the two homologous chromosomes of an individual. Consequently, only the unphased genotype, i.e. the combination of the two homologous haplotypes, is directly observable. Several methods have been proposed for inferring individual haplotypes and for estimating haplotype-specific relative risks based on unphased genotype data from case–control studies (see Schaid, 2004, for a recent review).

Cohort studies offer several advantages over case–control studies (Breslow and Day, 1987, pp. 11–20). First, the age of onset carries more information about the etiology of a complex disease than the disease status. Second, selection and information biases inherent in case–control studies can usually be eliminated in cohort studies. Third, the cohort design enables one to investigate a full range of diseases and related traits in a single study.

Cohort studies are major undertakings, involving long-term follow-up of many individuals. Fortunately, there are a number of cohort studies that have already been assembled for other purposes and have repositories of stored specimens that would allow the individuals to be genotyped for candidate genes of interest. Examples include the Cardiovascular Health Study (Fried *et al.*, 1991), the Women's Health Initiative (Johnson *et al.*, 1999), and the Atherosclerosis Risk in Communities (ARIC) Study (The ARIC Investigators, 1989).

Lin (2004) showed how to perform the Cox regression analysis of haplotype-disease associations with genotype data in cohort studies. The genotype data are required to be available on all cohort members. Despite the continuing improvement in genotyping efficiency, it is still prohibitively expensive to genotype a large cohort. An efficient compromise is to employ the case–cohort or nested case–control design (Kalbfleisch and Prentice, 2002, Section 11.4), so that only a subset of the cohort members need to be genotyped. In fact, the case–cohort design was recently employed in the ARIC study, which is an epidemiologic cohort study of 15 792 individuals aged 45–64 years to investigate the etiology of atherosclerosis and other diseases. There is a large body of literature on the Cox regression for case–cohort and nested case–control designs (see Kulich and Lin, 2004; Nan, 2004; Scheike and Juul, 2004; Scheike and Martinussen, 2004; and the references therein). None of the existing work, however, deals with the additional complexity due to haplotype uncertainty.

In the present paper, we study semiparametric estimation of haplotype-disease associations in case–cohort and nested case–control studies. The fact that the genotype data are available only on a biased subset of the cohort members poses considerable challenges in making inference about haplotype-disease associations. We propose a broad class of semiparametric regression models to formulate the effects of haplotype configurations and possibly time-dependent environment factors on the age of onset of disease. We derive appropriate likelihoods for these models and establish the asymptotic properties of the resultant maximum likelihood estimators. We develop efficient and stable numerical algorithms to implement the corresponding inference procedures. We apply the proposed methods to the aforementioned ARIC study, which motivated this work.

2. INFERENCE PROCEDURES

Let T be the time to disease occurrence, H the pair of homologous haplotypes, and G the corresponding genotype. If we denote the two possible alleles of each SNP by the values 0 versus 1, then H is a pair of ordered sequences of zeros and ones and G, which is the sum of the two sequences in H, is an ordered sequence of zeros, ones, and twos. Although we are interested in the association between H and T, we only observe G directly.

Under case–cohort and nested case–control designs, a subset of individuals is selected for genotyping. We allow the possibility that some other expensive discrete time-independent covariates, denoted by W, are also measured in this subset only. Additional covariates of interest X, possibly time dependent, are measured on all cohort members.

The time to disease occurrence will be censored if the individual has not developed the disease of interest by the end of the study or is withdrawn from the study prematurely. Let C denote the potential censoring time. We assume coarsening at random. That is, the event $C = t$ is independent of (T, H, W) conditional on $\{X(s): s \leqslant t\}$ and $T \geqslant t$.

Suppose that we have a cohort of n individuals. We collect the data $\{Y_i, \Delta_i, \overline{X}_i(Y_i)\}$ $(i = 1, \ldots, n)$, where $Y_i = \min(T_i, C_i)$, $\Delta_i = I(T_i \leqslant C_i)$, $\overline{X}_i(t) = \{X_i(s): s \leqslant t\}$, and $I(\cdot)$ is the indicator function. We also measure G and W for a subset of the cohort, which is selected by the case–cohort or nested case–control sampling.

Under the case–cohort sampling, we randomly select a subcohort from the full cohort. The selection probabilities depend on the observed event histories and possibly on covariates that are always measured. Let R_i indicate by the values 1 versus 0 whether the ith individual is selected. We assume missing at random in that $P(R_i = 1|T_i, C_i, H_i, W_i, X_i) = P(R_i = 1|Y_i, \Delta_i, \overline{X}_i(Y_i))$. The observed data can be represented as $(Y_i, \Delta_i, \overline{X}_i(Y_i), R_i, R_i G_i, R_i W_i)$ $(i = 1, \ldots, n)$.

Let $\mathcal{S}(G)$ denote the set of all haplotype pairs that are compatible with genotype G. Then the observed-data likelihood function can be written as

$$
\begin{aligned}
&\prod_{i=1}^{n} \Bigg[\sum_{H \in \mathcal{S}(G_i)} \lambda_T(Y_i|H, \overline{X}_i(Y_i), W_i)^{\Delta_i} \exp\left\{ -\int_0^{Y_i} \lambda_T(t|H, \overline{X}_i(t), W_i) \mathrm{d}t \right\} \\
&\qquad \times \prod_{t \leqslant Y_i} f_{X(t)}(X_i(t)|X_i(s), s < t, W_i, H) P(W_i|H) P(H) \Bigg]^{R_i} \\
&\times \Bigg[\sum_{W} \sum_{H} \lambda_T(Y_i|H, \overline{X}_i(Y_i), W)^{\Delta_i} \exp\left\{ -\int_0^{Y_i} \lambda_T(t|H, \overline{X}_i(t), W) \mathrm{d}t \right\} \\
&\qquad \times \prod_{t \leqslant Y_i} f_{X(t)}(X_i(t)|X_i(s), s < t, W, H) P(W|H) P(H) \Bigg]^{1-R_i} \\
&\times \lambda_C(Y_i|\overline{X}_i(Y_i))^{1-\Delta_i} \exp\left\{ -\int_0^{Y_i} \lambda_C(t|\overline{X}_i(t)) \mathrm{d}t \right\} \\
&\times P(R_i = 1|Y_i, \Delta_i, \overline{X}_i(Y_i))^{R_i} P(R_i = 0|Y_i, \Delta_i, \overline{X}_i(Y_i))^{1-R_i},
\end{aligned}
$$

where λ_T and λ_C pertain to the conditional hazard functions of T and C, respectively, and $f_{X(t)}$ pertains to the conditional density of $X(t)$. Thus, the observed-data likelihood function concerning the distribution

of T given (H, X, W) is proportional to

$$\prod_{i=1}^{n}\left[\sum_{H\in\mathcal{S}(G_i)}\lambda_T(Y_i|H,\overline{X}_i(Y_i),W_i)^{\Delta_i}\exp\left\{-\int_0^{Y_i}\lambda_T(t|H,\overline{X}_i(t),W_i)\mathrm{d}t\right\}\right.$$
$$\left.\times f(\overline{X}_i(Y_i)|W_i,H)P(W_i|H)P(H)\right]^{R_i}\left[\sum_{W}\sum_{H}\lambda_T(Y_i|H,\overline{X}_i(Y_i),W)^{\Delta_i}\right.$$
$$\left.\times\exp\left\{-\int_0^{Y_i}\lambda_T(t|H,\overline{X}_i(t),W)\mathrm{d}t\right\}\ f(\overline{X}_i(Y_i)|W,H)P(W|H)P(H)\right]^{1-R_i}, \qquad (2.1)$$

where f is the conditional density of $\overline{X}(t)$.

Under the nested case–control sampling, a small number of the individuals who are at risk at the time of disease occurrence of a case are selected for genotyping. The probability of selection at time t for an individual may depend on the observed past history $\mathcal{D}(t)$. The observed data can be represented as $\{Y_i, \Delta_i, \overline{X}_i(Y_i), \overline{S}_i(Y_i), \overline{S}_i(Y_i)G_i, \overline{S}_i(Y_i)W_i\}$ $(i = 1, \ldots, n)$, where $\overline{S}_i(t) = \{S_i(s)\colon s \leqslant t\}$ and $S_i(t)$ indicates whether the ith individual is selected for genotyping at time t.

To motivate the likelihood construction, we pretend that all the random variables are discrete. Then the observed-data likelihood is

$$\prod_{i=1}^{n}\left[\prod_t\{f_{X(t)}(X_i(t)|\mathcal{D}_i(t))P(T_i=t|\mathcal{D}_i(t))P(C_i>t|\mathcal{D}_i(t))P(W_i,G_i|T_i=t,\mathcal{D}_i(t))^{S_i(t)}\}^{\Delta_i I(Y_i=t)}\right.$$
$$\times\prod_t\{f_{X(t)}(X_i(t)|\mathcal{D}_i(t))P(C_i=t|\mathcal{D}_i(t))P(T_i>t|\mathcal{D}_i(t))P(W_i,G_i|T_i>t,\mathcal{D}_i(t))^{S_i(t)}\}^{(1-\Delta_i)I(Y_i=t)}$$
$$\times\prod_t\{P(C_i>t|\mathcal{D}_i(t))P(T_i>t|\mathcal{D}_i(t))P(W_i,G_i|T_i>t,\mathcal{D}_i(t))^{S_i(t)}\}^{I(Y_i>t)}$$
$$\left.\times\prod_t P(S_i(t)=1|\mathcal{D}_i(t),X_i(t))^{S_i(t)I(Y_i\geqslant t)}\{1-P(S_i(t)=1|\mathcal{D}_i(t),X_i(t))\}^{\{1-S_i(t)\}I(Y_i\geqslant t)}\right]. \qquad (2.2)$$

If the ith individual is never selected for genotyping, i.e. $S_i(t) = 0$ for all $t \leqslant Y_i$, then $\Delta_i = 0$ and $\mathcal{D}_i(t)$ only contains the information of $T_i \geqslant t$ and $\overline{X}_i(t)$, so the likelihood contribution from this individual is the same as the likelihood of $(Y_i, \Delta_i, \overline{X}_i(Y_i))$. If the ith individual is selected, i.e. $S_i(t) = 1$ for some $t_0 \leqslant Y_i$, then $\mathcal{D}_i(t)$ contains the information of $T_i \geqslant t$ and $\overline{X}_i(t)$ for $t < t_0$ and becomes the information of $T_i \geqslant t$, G_i, W_i, and $\overline{X}_i(t)$ for $t \geqslant t_0$, so the contribution from this individual to (2.2) is the same as the likelihood of $(Y_i, \Delta_i, G_i, W_i, \overline{X}_i(Y_i))$. Thus, the likelihood function concerning the distribution of T given (H, W, X) is exactly the same as (2.1), in which $R_i \equiv \max\{S_i(t)\colon t \leqslant Y_i\}$ indicates whether the ith individual is ever selected for genotyping.

REMARK 2.1 In the above derivation, the sampling is assumed to be independent among individuals. We may relax this assumption by allowing the sampling at time t to depend on the observed history at t of all individuals so that sampling without replacement can be accommodated. The likelihood function remains the same.

The conditional hazard function $\lambda_T(t|H, \overline{X}(t), W)$ represents the effects of the haplotype pair and environmental factors on the risk of disease, which can be formulated by a variety of parametric and

semiparametric models. We propose the following class of semiparametric transformation models in terms of the cumulative hazard function:

$$\Lambda_T(t|H, \overline{X}(t), W) = Q\left(\int_0^t \mathrm{e}^{\beta^T \mathcal{Z}(H, X(s), W)} \mathrm{d}\Lambda(s)\right), \tag{2.3}$$

where $\Lambda(t)$ is an unknown increasing function with $\Lambda(0) = 0$, $\mathcal{Z}(H, X(t), W)$ is a specified function of H, $X(t)$, and W, and Q is a three-time differentiable transformation with $Q(0) = 0$ and $Q'(x) > 0$. Here and in the sequel, $g'(x) = \mathrm{d}g(x)/\mathrm{d}x$ and $g''(x) = \mathrm{d}^2 g(x)/\mathrm{d}x^2$. We may use the class of Box–Cox transformations $Q(x) = \{(1 + x)^r - 1\}/r$ $(r > 0)$ or the class of logarithmic transformations $Q(x) = r_1 \log(1 + r_2 x)$ $(r_1 > 0, r_2 > 0)$. The choices of $Q(x) = x$ and $\log(1 + x)$ yield the proportional hazards and proportional odds models, respectively.

Nonidentifiability arises if the joint distribution of the haplotype pair is totally unrestricted. Lin (2004) assumed Hardy–Weinberg equilibrium such that $P(H = (h_k, h_l)) = \pi_k \pi_l$ $(k, l = 1, \ldots, K)$, where π_k is the marginal probability that the haplotype is h_k and K is the number of possible haplotypes. We consider the following one-parameter extension:

$$P(H = (h_k, h_l)) = \rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l, \quad k, l = 1, \ldots, K, \tag{2.4}$$

where $\delta_{kl} = 1$ if $k = l$ and 0 otherwise, and ρ is the inbreeding coefficient. Although the actual disequilibrium may not conform exactly to (2.4), this extension allows more robust inference than the standard Hardy–Weinberg equilibrium assumption.

Under (2.3) and (2.4), the observed-data likelihood function concerning the parameters of interest $\theta \equiv (\beta, \rho, \pi_1, \ldots, \pi_K)$ and Λ takes the form

$$\begin{aligned}
&\prod_{i=1}^{n} \Bigg[\sum_{H=(h_k,h_l)\in\mathcal{S}(G_i)} \left\{ \Lambda'(Y_i) \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(Y_i), W_i)} Q'\left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s), W_i)} \mathrm{d}\Lambda(s)\right)\right\}^{\Delta_i} \\
&\quad \times \exp\left\{ -Q\left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s), W_i)} \mathrm{d}\Lambda(s)\right)\right\} f(\overline{X}_i(Y_i)|W_i, H) P(W_i|H) \{\rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l\} \Bigg]^{R_i} \\
&\quad \times \Bigg[\sum_{W} \sum_{H=(h_k,h_l)} \left\{ \Lambda'(Y_i) \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(Y_i), W)} Q'\left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s), W)} \mathrm{d}\Lambda(s)\right)\right\}^{\Delta_i} \\
&\quad \times \exp\left\{ -Q\left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s), W)} \mathrm{d}\Lambda(s)\right)\right\} f(\overline{X}_i(Y_i)|W, H) P(W|H) \{\rho \pi_k \delta_{kl} + (1 - \rho)\pi_k \pi_l\} \Bigg]^{1-R_i}.
\end{aligned} \tag{2.5}$$

Simplifications arise under certain conditions. If there is no W, then (2.5) will not contain any term involving W. If $\overline{X}(t)$ is independent of (W, H), then the conditional density of $\overline{X}(Y)$ can be dropped out of (2.5) due to factorization. In the sequel, we focus on the most common situation in which W does not exist and X is independent of H.

We propose to estimate θ and Λ by the nonparametric maximum likelihood method. The maximum of (2.5) does not exist if Λ is restricted to be absolutely continuous. Thus, we allow Λ to be right continuous

and maximize the following function:

$$L_n(\theta, \Lambda) = \prod_{i=1}^{n} \left(\sum_{H=(h_k,h_l)\in\mathcal{S}(G_i)} \left[\Lambda\{Y_i\} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(Y_i))} Q' \left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s))} \mathrm{d}\Lambda(s) \right) \right]^{\Delta_i} \right.$$
$$\left. \times \exp\left\{ -Q\left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s))} \mathrm{d}\Lambda(s) \right) \right\} \{\rho\pi_k\delta_{kl} + (1-\rho)\pi_k\pi_l\} \right)^{R_i}$$
$$\times \left(\sum_{H=(h_k,h_l)} \left[\Lambda\{Y_i\} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(Y_i))} Q' \left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s))} \mathrm{d}\Lambda(s) \right) \right]^{\Delta_i} \right.$$
$$\left. \times \exp\left\{ -Q\left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(H, X_i(s))} \mathrm{d}\Lambda(s) \right) \right\} \{\rho\pi_k\delta_{kl} + (1-\rho)\pi_k\pi_l\} \right)^{1-R_i}, \qquad (2.6)$$

where $\Lambda\{Y_i\}$ denotes the jump size of Λ at Y_i. The maximization is tantamount to maximizing (2.6) over θ and the $\Lambda\{Y_i\}$ associated with $\Delta_i = 1$ and can be carried out through the EM algorithm described in Appendix A.

Let θ_0 and Λ_0 denote the true values of θ and Λ, and $\widehat{\theta}_n$ and $\widehat{\Lambda}_n$ the maximum likelihood estimators. We show in Appendix B that $n^{1/2}(\widehat{\theta}_n - \theta_0, \widehat{\Lambda}_n - \Lambda_0)$ weakly converges to a zero-mean Gaussian process and that the limiting covariance matrix of $n^{1/2}(\widehat{\theta}_n - \theta_0)$ achieves the semiparametric efficiency bound (Bickel *et al.*, 1993, Chapter 3). We can estimate the limiting covariance function of $n^{1/2}(\widehat{\theta}_n - \theta_0, \widehat{\Lambda}_n - \Lambda_0)$ by regarding (2.6) as a parametric likelihood with θ and the $\Lambda\{Y_i\}$ associated with $\Delta_i = 1$ as the parameters and inverting the observed information matrix for those parameters. We can also estimate the covariance matrix of $n^{1/2}(\widehat{\theta}_n - \theta_0)$ by the profile likelihood method (Murphy and van der Vaart, 2000). The profile log-likelihood function can be calculated via the EM algorithm, in which θ is held fixed.

3. Numerical results

3.1 *ARIC study*

We are currently evaluating common genetic polymorphisms which, in combination with exposure to tobacco smoking, may affect the risk of atherosclerosis and its clinical sequelae. An average of six polymorphisms, selected on the basis of their prevalence and functional significance, expression in relevant tissues, evaluation in previous studies, and biological plausibility within 19 genes involved in activation, detoxification, oxidative stress, and DNA repair pathways, are being evaluated in a well-characterized, bi-ethnic cohort of 15 792 men and women under active follow-up since 1987–1989 as part of the ARIC study. Four endpoints quantifying subclinical atherosclerosis and validated clinical atherosclerotic events are being studied under the case–cohort design.

So far, we have genotyped five SNPs in XRCC1, a major base excision repair gene. We considered all incident CHD cases occurring between 1987 and 2001. A subcohort was selected by stratified random sampling with different proportions of participants drawn from eight age–sex–race strata. Genotyping was conducted using matrix-assisted laser desorption/ionization time-of-flight mass spectrometry. Cigarette smoking history was obtained through an interviewer-administered questionnaire.

We focus on the Caucasian sample, which consists of 11 526 individuals, 774 cases, and a subcohort of 698 controls. Cigarette smoking status is known for 11 519 participants. The five SNPs are missing in 12%, 6%, 10%, 12%, and 6% of the case–cohort sample. The minor allele frequencies are 0.34, 0.40, 0.37,

0.41, and 0.36. There are nine haplotypes with estimated frequencies of higher than 0.5% in the sample. The frequencies for haplotypes (00100, 00110, 01001, 01100, 01110, 10110, 11001, 11100, 11110) are estimated at 0.012, 0.158, 0.096, 0.063, 0.012, 0.227, 0.276, 0.148, and 0.008, and the inbreeding coefficient is estimated at 0.025.

We fit separate models comparing each haplotype in turn with all others. Each model includes haplotype, smoking status (ever smoke $= 1$, never smoke $= 0$), two dummy variables contrasting Minnesota and Washington to North Carolina, gender and age at the baseline, as well as the interaction between smoking and haplotype. The effects of the haplotype pair are assumed to be additive (Lin, 2004). The results for the estimation of the haplotype effects and haplotype–smoking interactions under these models are summarized in Table 1. The individuals with haplotype 00110 appear to have a significantly higher risk of CHD as compared to the individuals without this haplotype. No estimate was obtained for haplotype 00100 due to numerical instability. There is no convincing evidence for interactions.

We also compare haplotype 00110 with the other five common haplotypes in a single model, and the estimation results are shown in Table 2. There is some evidence that haplotype 00110 is associated with higher risk of CHD than all other common haplotypes, especially haplotypes 01001, 10110, and 11001. The likelihood ratio statistic for testing the global null hypothesis of no haplotype effects and no haplotype–smoking interactions has an observed χ^2 value of 15.45 with 10 degrees of freedom, yielding a p-value of 0.116.

3.2 *Simulation studies*

We conducted extensive simulation studies to examine the finite-sample properties of the proposed methods. We considered five SNPs and generated genotypes according to the observed haplotype distribution of the ARIC data. We focused on the effect of haplotype 01100 and its interaction with a Bernoulli

Table 1. *Estimates of haplotype effects and haplotype–smoking interactions for the ARIC study based on separate models*

Haplotype	Parameter	Estimate	Standard error	p-value
00110	Main effect	0.237	0.105	0.024
	Interaction	−0.010	0.119	0.931
01001	Main effect	−0.295	0.239	0.218
	Interaction	0.003	0.273	0.992
01100	Main effect	0.124	0.243	0.610
	Interaction	−0.404	0.276	0.143
01110	Main effect	0.506	0.586	0.388
	Interaction	0.217	0.650	0.739
10110	Main effect	−0.078	0.143	0.585
	Interaction	0.102	0.171	0.551
11001	Main effect	0.165	0.146	0.259
	Interaction	0.048	0.177	0.786
11100	Main effect	0.029	0.166	0.863
	Interaction	−0.136	0.188	0.469
11110	Main effect	0.515	0.468	0.271
	Interaction	0.715	0.518	0.167

Note: Each haplotype is compared to all others. The analysis adjusts for geographical location, gender, and age.

Table 2. *Estimates of haplotype effects and haplotype–smoking interactions for the ARIC study based on a full model with haplotype 00110 as the reference*

Parameter	Estimate	Standard error	*p*-value
Haplotype 01001	−0.459	0.317	0.147
Haplotype 01100	−0.051	0.284	0.857
Haplotype 10110	−0.273	0.206	0.186
Haplotype 11001	−0.288	0.208	0.165
Haplotype 11100	−0.183	0.185	0.323
Smoking status	0.548	0.196	0.005
Minnesota	−0.129	0.077	0.093
Washington	0.214	0.071	0.003
Age	0.061	0.005	<0.001
Male	1.09	0.069	<0.001
01001× smoking	0.053	0.358	0.883
01100× smoking	−0.384	0.330	0.245
10110× smoking	0.052	0.235	0.826
11001× smoking	−0.018	0.231	0.936
11100× smoking	−0.103	0.205	0.616

environmental variable with 0.6 success probability, mimicking cigarette smoking in the ARIC data. We generated time to disease occurrence from either the proportional hazards model or the proportional odds model with baseline hazard function of $0.14t$ and with additive haplotype effects. The individuals were selected for genotyping by case–cohort or nested case–control sampling with two controls per case. The proportions of missingness for the five SNPs among those selected for genotyping were the same as in the ARIC study. We generated censoring times from the uniform $[0, 5]$ distribution truncated at 1. Approximately 90% of the observations were censored.

Table 3 summarizes the results of the simulation studies for $n = 2000$ and with various combinations of parameter values. The parameter estimators seem to have little bias. The profile likelihood method provides accurate estimators of the variances. The Wald tests have proper type I error rates, and the confidence intervals have reasonable coverage probabilities. The relative efficiencies of the case–cohort and nested case–control designs are generally between 80% and 90% for estimating haplotype effects and haplotype–environment interactions and over 95% for estimating environmental effects. Thus, these designs are highly cost-effective since only 30% of the entire cohort is genotyped. The case–cohort design appears to be slightly more efficient than the nested case–control design; however, 2–4% of selected controls became cases later on, so the total number of individuals genotyped is slightly smaller under the nested case–control design than under the case–cohort design.

4. Remarks

The results presented in Section 3.1 represent some preliminary findings from a major ongoing investigation. We are currently genotyping additional SNPs in the XRCC1 gene and examining 18 other genes using the methods proposed here. The full results will be reported elsewhere.

In practice, the true model is unknown. Thus, one will need to explore several possible models. Since the proposed methods are likelihood based, we can apply model selection criteria such as the Akaike information criterion (AIC) (Akaike, 1985) to determine the best model. Our experience shows that AIC performs well in this kind of setting (see Lin, 2004).

It is assumed in (2.6) that there is no W and X is independent of H. This assumption is reasonable in most genetic studies. It is easy to remove this assumption if X is time independent and discrete because

Table 3. *Summary statistics for the simulation studies*[†]

Parameter	Case–cohort design						Nested case–control design					
	Bias	SE	SEE	CP	PW	RE	Bias	SE	SEE	CP	PW	RE
						Proportional hazards model						
$\beta_h = 0$	−0.045	0.371	0.371	0.967	0.033	0.849	−0.042	0.379	0.373	0.960	0.040	0.803
$\beta_x = 0.5$	0.008	0.217	0.217	0.955	0.644	0.963	0.008	0.217	0.218	0.954	0.654	0.960
$\beta_{xh} = 0$	0.034	0.427	0.424	0.965	0.035	0.889	0.037	0.432	0.425	0.960	0.040	0.858
$\beta_h = 0.5$	−0.014	0.274	0.272	0.957	0.463	0.812	−0.011	0.280	0.274	0.954	0.463	0.779
$\beta_x = 0.5$	0.009	0.213	0.208	0.950	0.686	0.958	0.009	0.213	0.208	0.948	0.691	0.962
$\beta_{xh} = 0$	0.013	0.320	0.317	0.956	0.044	0.855	0.012	0.321	0.317	0.955	0.045	0.850
$\beta_h = 0$	−0.045	0.367	0.364	0.964	0.036	0.865	−0.041	0.371	0.365	0.962	0.038	0.844
$\beta_x = 0.5$	0.009	0.216	0.213	0.951	0.674	0.961	0.009	0.216	0.213	0.949	0.674	0.961
$\beta_{xh} = 0.5$	0.045	0.403	0.399	0.961	0.228	0.881	0.044	0.404	0.399	0.961	0.229	0.874
$\beta_h = 0.5$	−0.015	0.271	0.260	0.947	0.490	0.831	−0.014	0.277	0.261	0.944	0.490	0.793
$\beta_x = 0.5$	0.009	0.211	0.198	0.938	0.716	0.957	0.009	0.212	0.198	0.934	0.714	0.949
$\beta_{xh} = 0.5$	0.019	0.303	0.300	0.957	0.395	0.848	0.019	0.308	0.301	0.952	0.398	0.825
						Proportional odds model						
$\beta_h = 0$	−0.047	0.389	0.386	0.967	0.033	0.847	−0.043	0.389	0.389	0.962	0.038	0.830
$\beta_x = 0.5$	0.008	0.228	0.226	0.951	0.615	0.966	0.008	0.229	0.226	0.952	0.614	0.962
$\beta_{xh} = 0$	0.037	0.449	0.447	0.965	0.035	0.890	0.037	0.450	0.448	0.966	0.034	0.877
$\beta_h = 0.5$	−0.015	0.299	0.295	0.955	0.415	0.810	−0.013	0.307	0.297	0.948	0.419	0.769
$\beta_x = 0.5$	0.009	0.225	0.219	0.945	0.638	0.961	0.010	0.227	0.219	0.946	0.641	0.943
$\beta_{xh} = 0$	0.015	0.352	0.347	0.955	0.045	0.867	0.014	0.358	0.348	0.950	0.050	0.837
$\beta_h = 0$	−0.047	0.385	0.380	0.963	0.036	0.865	−0.045	0.392	0.382	0.961	0.039	0.830
$\beta_x = 0.5$	0.008	0.227	0.222	0.947	0.628	0.967	0.009	0.228	0.222	0.946	0.635	0.962
$\beta_{xh} = 0.5$	0.048	0.427	0.423	0.965	0.206	0.890	0.048	0.434	0.425	0.960	0.210	0.860
$\beta_h = 0.5$	−0.016	0.294	0.287	0.953	0.431	0.833	−0.014	0.301	0.289	0.947	0.427	0.800
$\beta_x = 0.5$	0.009	0.225	0.216	0.942	0.648	0.959	0.009	0.225	0.215	0.939	0.651	0.953
$\beta_{xh} = 0.5$	0.021	0.339	0.335	0.956	0.334	0.851	0.020	0.341	0.336	0.950	0.333	0.841

[†]β_h, β_x, and β_{xh} pertain to haplotype effect, environmental effect, and haplotype–environment interaction, respectively. Bias and SE denote the bias and standard error of parameter estimator, SEE is the mean of standard error estimator, CP is the coverage probability of 95% confidence interval, PW is the power for testing zero parameter, and RE is the efficiency relative to full-cohort design. Each entry is based on 5000 repetitions.

then the general likelihood function given in (2.5) just involves some discrete probability functions. If X contains one or two time-independent continuous components, we can still estimate the conditional density function of X nonparametrically.

We have assumed that W is discrete and time independent. If W is continuous and possibly time dependent but X is discrete and time independent, we will replace $f(\overline{X}(Y)|W, H)P(W|H)$ in (2.5) with $f(\overline{W}(Y)|X, H)P(X|H)$. However, if both X and W are continuous, it is necessary to parameterize the distribution; nevertheless, nonparametric estimation is possible if X and W have one continuous component each.

If one is not interested in haplotypes, the likelihood given in (2.5) simplifies greatly. There will be no summation over H, and H will disappear from all expressions. The theoretical results will continue to hold, and the EM algorithm will still apply, although the parameters will not include ρ and π_k. Scheike

and Juul (2004) and Scheike and Martinussen (2004) studied maximum likelihood estimation in the proportional hazards model under case–cohort and nested case–control designs (without the additional complexities due to haplotype uncertainty and missing genotype data) but did not provide theoretical justifications for the asymptotic results. The asymptotic theory derived in the present paper covers those situations. Note that the aforementioned challenge in dealing with continuous covariates still exists even when one is not interested in haplotypes. In fact, this challenge tends to be less severe in genetic studies because genes are discrete and are usually independent of other covariates.

This paper is focused on case–cohort and nested case–control designs, while the recent paper of Lin and Zeng (2006) is concerned with other commonly used study designs. A nontechnical description of the methods developed in the two papers was provided by Lin *et al.* (2005). The software is available at http://www.bios.unc.edu/∼lin.

ACKNOWLEDGMENTS

This research was supported by the National Institutes of Health. The authors thank two referees for their timely reviews and helpful comments. *Conflict of Interest*: None declared.

APPENDIX A

EM algorithm

Write $H = BQ_1 + (1 - B)Q_2$, where B is a Bernoulli variable with success probability ρ and Q_1 and Q_2 are discrete variables with $P(Q_1 = (h_k, h_k)) = \pi_k$ and $P(Q_2 = (h_k, h_l)) = \pi_k\pi_l$ $(k, l = 1, \ldots, K)$. We introduce a subject-specific frailty ξ with density $\phi(\xi)$ such that

$$e^{-Q(x)} = \int_0^\infty e^{-xt}\phi(t)dt. \tag{A.1}$$

Then the observed-data likelihood function under the transformation model is equivalent to the likelihood function under the proportional hazards frailty model: the conditional hazard function of T given (B, Q_1, Q_2, ξ) is $\lambda(t)\xi \exp\{\beta^T Z(BQ_1 + (1-B)Q_2, X(t))\}$. By treating (B, Q_1, Q_2, ξ) as missing data, we obtain the following complete-data likelihood function:

$$\prod_{i=1}^n \left[\Lambda\{Y_i\}\xi_i e^{\beta^T Z(B_i Q_{1i}+(1-B_i)Q_{2i}, X_i(Y_i))}\right]^{\Delta_i} \exp\left\{-\xi_i \int_0^{Y_i} e^{\beta^T Z(B_i Q_{1i}+(1-B_i)Q_{2i}, X_i(s))} d\Lambda(s)\right\}$$

$$\times \rho^{B_i}(1-\rho)^{1-B_i} \prod_{k=1}^K \pi_k^{B_i I(Q_{1i}=(h_k,h_k))} \prod_{k,l=1}^K (\pi_k\pi_l)^{(1-B_i)I(Q_{2i}=(h_k,h_l))}. \tag{A.2}$$

In the M-step of the EM algorithm, we maximize the conditional expectation of the logarithm of (A.2) given the observed data. Let $\widehat{E}_i[\cdot]$ denote the conditional expectation given the ith observation $(Y_i, X_i, \Delta_i, R_i, R_i G_i)$. Then ρ and π_k are updated by the following formulas:

$$\rho = n^{-1} \sum_{i=1}^n \widehat{E}_i[B_i],$$

$$\pi_k = n^{-1} \sum_{i=1}^n \widehat{E}_i \left[B_i I(Q_{1i} = (h_k, h_k)) + 2 \sum_{l=1}^K (1 - B_i) I(Q_{2i} = (h_k, h_l)) \right].$$

In addition, we update β by solving the following equation:

$$\sum_{i=1}^{n} \Delta_i \left(\widehat{\mathrm{E}}_i[\mathcal{Z}(H_i, X_i(Y_i))] - \frac{\sum_{j=1}^{n} I(Y_j \geqslant Y_i) \widehat{\mathrm{E}}_j \left[\xi_j \mathcal{Z}(H_j, X_j(Y_i)) \mathrm{e}^{\beta^T \mathcal{Z}(H_j, X_j(Y_i))} \right]}{\sum_{j=1}^{n} I(Y_j \geqslant Y_i) \widehat{\mathrm{E}}_j \left[\xi_j \mathrm{e}^{\beta^T \mathcal{Z}(H_j, X_j(Y_i))} \right]} \right) = 0, \quad \text{(A.3)}$$

and update Λ by the step function with jump sizes

$$\Lambda\{Y_i\} = \Delta_i \Big/ \sum_{j=1}^{n} I(Y_j \geqslant Y_i) \widehat{\mathrm{E}}_j \left[\xi_j \mathrm{e}^{\beta^T \mathcal{Z}(H_j, X_j(Y_i))} \right], \quad i = 1, \ldots, n. \quad \text{(A.4)}$$

Note that (A.3) and (A.4) are reminiscent of the partial likelihood (Cox, 1972) score equation and the Breslow (1972) estimator.

In light of (A.3) and (A.4), we calculate the conditional expectations in the form of $\mathrm{E}[\xi_i \omega(B_i, Q_{1i}, Q_{2i}) | Y_i, X_i, \Delta_i, R_i, R_i G_i]$ in the E-step. We can avoid numerical integration over ξ_i in these calculations. Define $U_i(b, q_1, q_2) = \int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(bq_1 + (1-b)q_2, X_i(s))} \mathrm{d}\Lambda(s)$. In view of (A.2), the conditional density of ξ given (B_i, Q_{1i}, Q_{2i}) and $(Y_i, X_i, \Delta_i, R_i, R_i G_i)$ is proportional to $\xi^{\Delta_i} \mathrm{e}^{-\xi U_i(B_i, Q_{1i}, Q_{2i})} \phi(\xi)$, so that

$$\begin{aligned} &\mathrm{E}[\xi_i | B_i, Q_{1i}, Q_{2i}, Y_i, X_i, \Delta_i, R_i, R_i G_i] \\ &\quad = \int \xi^{1+\Delta_i} \mathrm{e}^{-\xi U_i(B_i, Q_{1i}, Q_{2i})} \phi(\xi) \mathrm{d}\xi \Big/ \int \xi^{\Delta_i} \mathrm{e}^{-\xi U_i(B_i, Q_{1i}, Q_{2i})} \phi(\xi) \mathrm{d}\xi. \end{aligned}$$

By differentiating (A.1) with respect to x, we obtain

$$\mathrm{e}^{-Q(x)} Q'(x) = \int \xi \mathrm{e}^{-\xi x} \phi(\xi) \mathrm{d}\xi, \quad -\mathrm{e}^{-Q(x)} \{Q''(x) - Q'(x)^2\} = \int \xi^2 \mathrm{e}^{-\xi x} \phi(\xi) \mathrm{d}\xi.$$

It follows that

$$\mathrm{E}[\xi_i | B_i, Q_{1i}, Q_{2i}, Y_i, X_i, \Delta_i, R_i, R_i G_i] = Q'(U_i(B_i, Q_{1i}, Q_{2i})) - \Delta_i \frac{Q''(U_i(B_i, Q_{1i}, Q_{2i}))}{Q'(U_i(B_i, Q_{1i}, Q_{2i}))}.$$

Consequently,

$$\begin{aligned} \mathrm{E}[\xi_i \omega(B_i, Q_{1i}, Q_{2i}) | Y_i, X_i, \Delta_i, R_i, R_i G_i] = \mathrm{E} \Bigg[\Bigg\{ Q'(U_i(B_i, Q_{1i}, Q_{2i})) - \Delta_i \frac{Q''(U_i(B_i, Q_{1i}, Q_{2i}))}{Q'(U_i(B_i, Q_{1i}, Q_{2i}))} \Bigg\} \\ \times \omega(B_i, Q_{1i}, Q_{2i}) \Big| Y_i, X_i, \Delta_i, R_i, R_i G_i \Bigg]. \end{aligned}$$

According to (A.2), the conditional density of (B_i, Q_{1i}, Q_{2i}) given the observed data is proportional to $g_i(B_i, Q_{1i}, Q_{2i})$, where

$$\begin{aligned} g_i(b, q_1, q_2) = &\left\{ \mathrm{e}^{\beta^T \mathcal{Z}(bq_1 + (1-b)q_2, X_i(Y_i))} Q' \left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(bq_1 + (1-b)q_2, X_i(s))} \mathrm{d}\Lambda(s) \right) \right\}^{\Delta_i} \\ &\times \exp \left\{ -Q \left(\int_0^{Y_i} \mathrm{e}^{\beta^T \mathcal{Z}(bq_1 + (1-b)q_2, X_i(s))} \mathrm{d}\Lambda(s) \right) \right\} \\ &\times \rho^b (1-\rho)^{1-b} \prod_{k=1}^{K} \pi_k^{bI(q_1 = (h_k, h_k))} \prod_{k,l=1}^{K} (\pi_k \pi_l)^{(1-b) I(q_2 = (h_k, h_l))}. \end{aligned}$$

Thus, $\mathrm{E}[\xi_i \omega(B_i, Q_{1i}, Q_{2i})|Y_i, X_i, \Delta_i, R_i, R_i G_i]$ is equal to

$$\frac{\sum_{bq_1+(1-b)q_2\in\mathcal{S}(G_i)} g_i(b,q_1,q_2)\omega(b,q_1,q_2)\{Q'(U_i(b,q_1,q_2))-\Delta_i Q''(U_i(b,q_1,q_2))/Q'(U_i(b,q_1,q_2))\}}{\sum_{bq_1+(1-b)q_2\in\mathcal{S}(G_i)} g_i(b,q_1,q_2)}$$

for individuals with $R_i = 1$ and is equal to

$$\frac{\sum_{b,q_1,q_2} g_i(b,q_1,q_2)\omega(b,q_1,q_2)\{Q'(U_i(b,q_1,q_2)) - \Delta_i Q''(U_i(b,q_1,q_2))/Q'(U_i(b,q_1,q_2))\}}{\sum_{b,q_1,q_2} g_i(b,q_1,q_2)}$$

for individuals with $R_i = 0$.

APPENDIX B

Asymptotic results

We impose the following conditions:

(C.1) Both $X(t)$ and $\mathcal{Z}(H, X(t))$ have bounded total variations in $[0, \tau]$ with probability one, where τ corresponds to the end of the study.
(C.2) There exists a positive constant a such that with probability one, $P(R = 1|Y, \Delta, \overline{X}(Y)) > a$ and $P(C \geqslant \tau|\overline{X}(\tau)) = P(C = \tau|\overline{X}(\tau)) > a$.
(C.3) If $\mu_1(t)+\beta_1^T \mathcal{Z}((h_k, h_k), X(t)) = \mu_2(t)+\beta_2^T \mathcal{Z}((h_k, h_k), X(t))$ for $t \in [0, \tau]$ and $k = 1, \ldots, K$ with probability one, then $\beta_1 = \beta_2$ and $\mu_1(t) = \mu_2(t)$.
(C.4) $|\beta_0| \leqslant c_0$ for some known constant c_0, and $\lambda_0(t)$ is continuous and positive for $t \in [0, \tau]$.
(C.5) $Q(x)$ satisfies one of the two conditions:

(C.5.1) for any positive constant c_0, $\limsup_{x\to\infty} \{Q(c_0 x)\}^{-1} \log\{x \sup_{y\leqslant x} Q'(y)\} = 0$,
(C.5.2) there exist some constants $r_1, r_2 > 0$ such that $Q(x) = r_1 \log(1 + r_2 x)$.

We state the asymptotic results in three theorems. The above conditions are assumed to hold in the theorems. The first theorem states the consistency, weak convergence, and asymptotic efficiency.

THEOREM B.1 With probability one,

$$|\widehat{\theta}_n - \theta_0| + \sup_{t\in[0,\tau]} |\widehat{\Lambda}_n(t) - \Lambda_0(t)| \to 0.$$

In addition, $n^{1/2}(\widehat{\theta}_n - \theta_0, \widehat{\Lambda}_n - \Lambda_0)$ weakly converges to a zero-mean Gaussian process in $R^d \times l^\infty[0, \tau]$, where d is the dimension of θ and $l^\infty[0, \tau]$ is a normed space consisting of all the bounded functions and the norm is defined as the supremum norm on $[0, \tau]$. Furthermore, the limiting covariance matrix of $n^{1/2}(\widehat{\theta}_n - \theta_0)$ achieves the semiparametric efficiency bound.

The second theorem justifies the estimation of the limiting covariance function of $n^{1/2}(\widehat{\theta}_n - \theta_0, \widehat{\Lambda}_n - \Lambda_0)$ by the inverse information matrix.

THEOREM B.2 Let $V(h_1, h_2)$ be the limiting variance of the random variable $n^{1/2}\big[h_1^T(\widehat{\theta}_n - \theta_0) + \int_0^\tau h_2(t)\mathrm{d}\{\widehat{\Lambda}_n(t) - \Lambda_0(t)\}\big]$, where h_1 is a d-vector and h_2 is a bounded function. The estimator $n h_n^T \mathcal{I}_n^{-1} h_n \to V(h_1, h_2)$ uniformly in (h_1, h_2) in probability, where h_n consists of h_1 and the values of $h_2(Y_i)$ associated with $\Delta_i = 1$, and $\mathcal{I}_n$ is the negative Hessian matrix of $\log L_n(\widehat{\theta}_n, \widehat{\Lambda}_n)$ with respect to θ and the $\Lambda\{Y_i\}$ associated with $\Delta_i = 1$.

The last theorem justifies the use of the profile log-likelihood $pl_n(\theta) \equiv \max_\Lambda \log L_n(\theta, \Lambda)$ in estimating the limiting covariance matrix of $n^{1/2}(\widehat{\theta}_n - \theta_0)$.

THEOREM B.3 For any d-vector h_1 with norm one,

$$-\frac{pl_n(\widehat{\theta}_n + \epsilon_n h_1) - 2pl_n(\widehat{\theta}_n) + pl_n(\widehat{\theta}_n - \epsilon_n h_1)}{n\epsilon_n^2} \to h_1^T \Omega^{-1} h_1$$

in probability, where $\epsilon_n = \mathrm{O}(n^{-1/2})$ and Ω is the limiting covariance matrix of $n^{1/2}(\widehat{\theta}_n - \theta_0)$.

The proofs of these theorems involve advanced mathematical tools from empirical process theory (van der Vaart and Wellner, 1996) and semiparametric efficiency theory (Bickel *et al.*, 1993). We outline here the main arguments. The detailed proofs are available from the authors.

Proof of Theorem B.1. We first prove the consistency under Condition (C.5.1). The proof consists of three steps.

Step 1: We show the existence of $(\widehat{\theta}_n, \widehat{\Lambda}_n)$ or equivalently the finiteness of the jump sizes of $\widehat{\Lambda}_n$. The logarithm of (2.6), denoted by $l_n(\theta, \Lambda)$, is bounded by

$$\mathrm{O}(1) + \sum_{i=1}^{n} \left(\Delta_i \log \left[\Lambda\{Y_i\} \sup_{y \leqslant \mathrm{e}^M \Lambda(Y_i)} Q'(y) \right] - Q(\mathrm{e}^{-M} \Lambda(Y_i)) \right), \tag{B.1}$$

where $\mathrm{O}(1)$ denotes some positive constant and M is a constant satisfying

$$\mathrm{e}^{-M} \leqslant \inf_{t,\beta,H,X} \exp\{\beta^T \mathcal{Z}(H, X(t))\} \leqslant \sup_{t,\beta,H,X} \exp\{\beta^T \mathcal{Z}(H, X(t))\} \leqslant \mathrm{e}^M.$$

Such an M exists under Conditions (C.1) and (C.4). It then follows from Condition (C.5.1) that (B.1) will diverge if $\Lambda\{Y_i\}$ is infinite for some i.

Step 2: We show that with probability one, $\widehat{\Lambda}_n$ is bounded for any n. Let $\overline{\Lambda}_n = \widehat{\Lambda}_n / \psi_n$, where $\psi_n = \widehat{\Lambda}_n(\tau)$. Clearly,

$$\begin{aligned} 0 \leqslant n^{-1}\{l_n(\widehat{\theta}_n, \widehat{\Lambda}_n) - l_n(\widehat{\theta}_n, \overline{\Lambda}_n)\} \leqslant \mathrm{O}(1) + n^{-1} \sum_{i=1}^{n} \Delta_i \log \left\{ \psi_n \sup_{y \leqslant \mathrm{e}^M \psi_n} Q'(y) \right\} \\ - n^{-1} \sum_{i=1}^{n} (1 - \Delta_i) I(Y_i = \tau) Q(\mathrm{e}^{-M} \psi_n). \end{aligned} \tag{B.2}$$

Since $P(\Delta = 0, Y = \tau) > 0$, (B.2) will be negative if ψ_n diverges. Thus, ψ_n is bounded, which implies that $\widehat{\Lambda}_n$ is bounded.

Step 3: By Helly's selection theorem, we can choose a subsequence such that $\widehat{\theta}_n \to \theta^*$ and $\widehat{\Lambda}_n \to \Lambda^*$ with probability one. It remains to show that $\theta^* = \theta_0$ and $\Lambda^* = \Lambda_0$. Note that

$$\widehat{\Lambda}_n\{Y_i\} = \Delta_i / \{n\phi_n(Y_i; \widehat{\theta}_n, \widehat{\Lambda}_n)\}, \tag{B.3}$$

where

$$\phi_n(t;\theta,\Lambda) = n^{-1}\sum_{i=1}^{n}\frac{R_i\sum_{H\in\mathcal{S}(G_i)}D_{1i}(\theta,\Lambda)D_{2i}(t;\theta,\Lambda)}{\sum_{H\in\mathcal{S}(G_i)}D_{1i}(\theta,\Lambda)} + n^{-1}\sum_{i=1}^{n}\frac{(1-R_i)\sum_{H}D_{1i}(\theta,\Lambda)D_{2i}(t;\theta,\Lambda)}{\sum_{H}D_{1i}(\theta,\Lambda)},$$

$$D_{1i}(\theta,\Lambda) = \left\{e^{\beta^T Z(H,X_i(Y_i))}Q'\left(\int_0^{Y_i}e^{\beta^T Z(H,X_i(s))}d\Lambda(s)\right)\right\}^{\Delta_i} \times \exp\left\{-Q\left(\int_0^{Y_i}e^{\beta^T Z(H,X_i(s))}d\Lambda(s)\right)\right\}\{\rho\pi_k\delta_{kl}+(1-\rho)\pi_k\pi_l\},$$

and

$$D_{2i}(t;\theta,\Lambda) = \frac{\Delta_i Q''\left(\int_0^{Y_i}e^{\beta^T Z(H,X_i(s))}d\Lambda(s)\right)e^{\beta^T Z(H,X_i(t))}I(Y_i\geqslant t)}{Q'\left(\int_0^{Y_i}e^{\beta^T Z(H,X_i(s))}d\Lambda(s)\right)} - Q'\left(\int_0^{Y_i}e^{\beta^T Z(H,X_i(s))}d\Lambda(s)\right)e^{\beta^T Z(H,X_i(t))}I(Y_i\geqslant t).$$

In view of (B.3), we construct another step function $\widetilde{\Lambda}_n$ with $\widetilde{\Lambda}_n\{Y_i\} = \Delta_i/\{n\phi_n(Y_i;\theta_0,\Lambda_0)\}$. By the Glivenko–Cantelli theorem, $\widetilde{\Lambda}_n$ uniformly converges to Λ_0, and $\widehat{\Lambda}_n$ is absolutely continuous with respect to $\widetilde{\Lambda}_n$ with the derivative converging uniformly to $d\Lambda^*(t)/d\Lambda_0(t)$. Since $n^{-1}\{l_n(\widehat{\theta}_n,\widehat{\Lambda}_n) - l_n(\theta_0,\widetilde{\Lambda}_n)\}\geqslant 0$, the Kullback–Leibler information of (θ^*,Λ^*) with respect to (θ_0,Λ_0) is non-negative, so that (2.6) has the same value almost surely whether $(\theta,\Lambda) = (\theta^*,\Lambda^*)$ or (θ_0,Λ_0). Setting $n=1$, $G_i = 2h_k$, $R_i = 1$, and $\Delta_i = 1$ and integrating Y_i from y to τ, we obtain

$$\left[\exp\left\{-Q\left(\int_0^{y}e^{\beta^{*T}Z((h_k,h_k),X(s))}d\Lambda^*(s)\right)\right\} - \exp\left\{-Q\left(\int_0^{\tau}e^{\beta^{*T}Z((h_k,h_k),X(s))}d\Lambda^*(s)\right)\right\}\right] \times\left\{\rho^*\pi_k^* + (1-\rho^*)\pi_k^{*2}\right\} = \left[\exp\left\{-Q\left(\int_0^{y}e^{\beta_0{}^TZ((h_k,h_k),X(s))}d\Lambda_0(s)\right)\right\} - \exp\left\{-Q\left(\int_0^{\tau}e^{\beta_0{}^TZ((h_k,h_k),X(s))}d\Lambda_0(s)\right)\right\}\right]\{\rho_0\pi_{0k}+(1-\rho_0)\pi_{0k}^2\}.$$

By comparing this equation with the one obtained from (2.6) with $n=1$, $G_i = 2h_k$, $R_i = 1$, $\Delta_i = 0$, and $Y_i = \tau$, we have

$$\exp\left\{-Q\left(\int_0^{y}e^{\beta^{*T}Z((h_k,h_k),X(s))}d\Lambda^*(s)\right)\right\}\left\{\rho^*\pi_k^* + (1-\rho^*)\pi_k^{*2}\right\} = \exp\left\{-Q\left(\int_0^{y}e^{\beta_0{}^TZ((h_k,h_k),X(s))}d\Lambda_0(s)\right)\right\}\{\rho_0\pi_{0k}+(1-\rho_0)\pi_{0k}^2\}.$$

The choice of $y = 0$ yields that $\rho^* \pi_k^* + (1 - \rho^*)\pi_k^{*2} = \rho_0 \pi_{0k} + (1 - \rho_0)\pi_{0k}^2$, which entails that $\rho^* = \rho_0$ and $\pi_k^* = \pi_{0k}$. In addition, $\int_0^y \mathrm{e}^{\beta^{*T} \mathcal{Z}((h_k,h_k),X(s))} \mathrm{d}\Lambda^*(s) = \int_0^y \mathrm{e}^{\beta_0{}^T \mathcal{Z}((h_k,h_k),X(s))} \mathrm{d}\Lambda_0(s)$, implying that

$$\log \lambda^*(y) + \beta^{*T} \mathcal{Z}((h_k, h_k), X(y)) = \log \lambda_0(y) + \beta_0^T \mathcal{Z}((h_k, h_k), X(y)).$$

It then follows from Condition (C.6) that $\beta^* = \beta_0$ and $\Lambda^* = \Lambda_0$. Hence, $\widehat{\theta}_n \to \theta_0$ and $\widehat{\Lambda}_n \to \Lambda_0$ almost surely. Since Λ_0 is continuous, the weak convergence of $\widehat{\Lambda}_n$ can be strengthened to the convergence uniformly in $[0, \tau]$.

If $Q(\cdot)$ satisfies Condition (C.5.2) instead of (C.5.1), we need to modify Step 2. It follows from (B.3) that

$$n\widehat{\Lambda}_n\{Y_i\} \geqslant \mathrm{O}(1) n^{-1} \sum_{k=1}^{n} \frac{I(Y_k \geqslant Y_i)}{1 + r_2 \mathrm{e}^M \widehat{\Lambda}_n(Y_k)}.$$

Thus,

$$\begin{aligned} 0 &\leqslant n^{-1}\{l_n(\widehat{\theta}_n, \widehat{\Lambda}_n) - l_n(\theta_0, \Lambda_0)\} \\ &\leqslant \mathrm{O}(1) - \mathrm{O}(1) n^{-1} \sum_{i=1}^{n} \log\{1 + r_2 \mathrm{e}^{-M} \widehat{\Lambda}_n(Y_i)\} - n^{-1} \sum_{i=1}^{n} \Delta_i \log \left\{ \mathrm{O}(1) n^{-1} \sum_{k=1}^{n} \frac{I(Y_k \geqslant Y_i)}{1 + r_2 \mathrm{e}^M \widehat{\Lambda}_n(Y_k)} \right\}. \end{aligned} \tag{B.4}$$

By partitioning $[0, \tau]$ into a sequence of intervals as in Zeng *et al.* (2005) and examining the two terms on the right-hand side of (B.4) when Y_i lies in each partition, we can show that the right-hand side of (B.4) is negative if $\widehat{\Lambda}_n(\tau)$ diverges. Thus, $\widehat{\Lambda}_n$ must be bounded.

To derive the asymptotic distribution of $(\widehat{\theta}_n, \widehat{\Lambda}_n)$, we apply Theorem 3.3.1 of van der Vaart and Wellner (1996) to the score operators for $\widehat{\theta}_n$ and $\widehat{\Lambda}_n$. Except for the invertibility of the derivative operator of the score operator, all the conditions in Theorem 3.3.1 can be verified via empirical process theory (see Zeng *et al.*, 2005). The derivative operator is invertible if the information operator is one-to-one. Thus, we wish to show that if a score function along the path $(\theta_0 + \epsilon h_1, \Lambda_0 + \epsilon \int h_2(t) \mathrm{d}\Lambda_0)$ is zero, then $h_1 = 0$ and $h_2 = 0$. For $R_i = 1$ and $G_i = 2h_k$, the score equation is

$$\begin{aligned} &h_2(Y) + h_{1\beta}^T \mathcal{Z}((h_k, h_k), X(Y)) \\ &+ \left\{ \frac{Q''\left(\int_0^Y \mathrm{e}^{\beta_0^T \mathcal{Z}((h_k,h_k),X(s))} \mathrm{d}\Lambda_0(s)\right)}{Q'\left(\int_0^Y \mathrm{e}^{\beta_0^T \mathcal{Z}((h_k,h_k),X(s))} \mathrm{d}\Lambda_0(s)\right)} - Q'\left(\int_0^Y \mathrm{e}^{\beta_0^T \mathcal{Z}((h_k,h_k),X(s))} \mathrm{d}\Lambda_0(s)\right) \right\} \\ &\times \left[\int_0^Y \mathrm{e}^{\beta_0^T \mathcal{Z}((h_k,h_k),X(s))} \{h_2(s) + h_{1\beta}^T \mathcal{Z}((h_k, h_k), X(s))\} \mathrm{d}\Lambda_0(s) \right] \\ &+ \{h_{1\rho}(\pi_{0k} - \pi_{0k}^2) + h_{1k}(\rho_0 + 2(1 - \rho_0)\pi_{0k})\}/(\rho_0 \pi_{0k} + (1 - \rho_0)\pi_{0k}^2) = 0, \end{aligned} \tag{B.5}$$

where $(h_{1\beta}, h_{1\rho}, h_{1k})$ are the components of h_1 associated with $(\beta_0, \rho_0, \pi_{0k})$ and $\sum_k h_{1k} = 0$. Setting $Y = 0$ yields that $h_{1\rho} = h_{1k} = 0$. This result implies that (B.5) is a homogeneous integral equation for $h_2(Y) + h_\beta^T \mathcal{Z}((h_k, h_k), X(Y))$, so that $h_2(Y) + h_\beta^T \mathcal{Z}((h_k, h_k), X(Y)) = 0$. Thus, $h_2 = 0$ and $h_{1\beta} = 0$. It then follows from Theorem 3.3.1 of van der Vaart and Wellner (1996) that $n^{1/2}(\widehat{\theta}_n - \theta_0, \widehat{\Lambda}_n - \Lambda_0)$ weakly converges to a zero-mean Gaussian process. Furthermore, we can use the arguments of Zeng *et al.* (2005) to show that $\widehat{\theta}_n$ is asymptotically efficient. □

Proof of Theorem B.2. This proof follows from the arguments given in the proof of Theorem 3 of Zeng *et al*. (2005). The details are omitted. □

Proof of Theorem B.3. We can verify the conditions in Theorem 1 of Murphy and van der Vaart (2000). In particular, we can construct the least favorable submodel by using the invertibility of the information operator shown in the proof of Theorem 1. The details are omitted. □

REFERENCES

AKAIKE, H. (1985). Prediction and entropy. In Atkinson, A. C. and Fienberg, S. E. (eds), *A Celebration of Statistics*. New York: Springer, pp. 1–24.

AKEY, J., JIN, L. AND XIONG, M. (2001). Haplotypes vs. single marker linkage disequilibrium tests: what do we gain? *European Journal of Human Genetics* **9**, 291–300.

BICKEL, P. J., KLASSEN, C. A. J., RITOV, Y. AND WELLNER, J. A. (1993). *Efficient and Adaptive Estimation in Semiparametric Models*. Baltimore, MD: Johns Hopkins University Press.

BRESLOW, N. E. (1972). Discussion of the paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* **34**, 216–217.

BRESLOW, N. E. AND DAY, N. E. (1987). *Statistical Methods in Cancer Research: The Design and Analysis of Cohort Studies.* Lyon: International Agency for Research on Cancer.

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

FRIED, L. P., BORHANI, N. O., ENRIGHT, P., FURBERG, C. D., GARDIN, J. M., KRONMAL, R. A., KULLER, L. H., MANOLIO, T. A., MITTELMARK, M. B., NEWMAN, A. *et al.* (1991). The Cardiovascular Health Study: design and rationale. *Annals of Epidemiology* **1**, 263–276.

JOHNSON, S. R., ANDERSON, G. L., BARAD, D. H. AND STEFANICK, M. L. (1999). The Women's Health Initiative: rationale, design, and progress report. *Journal of the British Menopause Society* **5**, 155–159.

KALBFLEISCH, J. D. AND PRENTICE, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. Hoboken, NJ: Wiley.

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

LIN, D. Y. (2004). Haplotype-based association analysis in cohort studies of unrelated individuals. *Genetic Epidemiology* **26**, 255–264.

LIN, D. Y. AND ZENG, D. (2006). Likelihood-based inference on haplotype effects in genetic association studies (with discussion). *Journal of the American Statistical Association* **101**, 89–118.

LIN, D. Y., ZENG, D. AND MILLIKAN, R. (2005). Maximum likelihood estimation of haplotype effects and haplotype-environment interactions in association studies. *Genetic Epidemiology* **29**, 299–312.

MORRIS, R. W. AND KAPLAN, N. L. (2002). On the advantage of haplotype analysis in the presence of multiple disease susceptibility alleles. *Genetic Epidemiology* **23**, 221–233.

MURPHY, S. A. AND VAN DER VAART, A. W. (2000). On profile likelihood. *Journal of the American Statistical Association* **95**, 449–465.

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

SCHAID, D. J. (2004). Evaluating associations of haplotypes with traits. *Genetic Epidemiology* **27**, 348–364.

SCHAID, D. J., ROWLAND, C. M., TINES, D. E., JACOBSON, R. M. AND POLAND, G. A. (2002). Score tests for association between traits and haplotypes when linkage phase is ambiguous. *American Journal of Human Genetics* **70**, 425–434.

SCHEIKE, T. H. AND JUUL, A. (2004). Maximum likelihood estimation for Cox's regression model under nested case-control sampling. *Biostatistics* **5**, 193–206.

SCHEIKE, T. H. AND MARTINUSSEN, T. (2004). Maximum likelihood estimation for Cox's regression model under case-cohort sampling. *Scandinavian Journal of Statistics* **31**, 283–293.

THE ARIC INVESTIGATORS (1989). The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. *American Journal of Epidemiology* **129**, 687–702.

VAN DER VAART, A. W. AND WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes*. New York: Springer.

ZAYKIN, D. V., WESTFALL, P. H., YOUNG, S. S., KARNOUB, M. A., WAGNER, M. J. AND EHM, M. G. (2002). Testing association of statistically inferred haplotypes with discrete and continuous traits in samples of unrelated individuals. *Human Heredity* **53**, 79–91.

ZENG, D., LIN, D. Y. AND YIN, G. (2005). Maximum likelihood estimation for the proportional odds model with random effects. *Journal of the American Statistical Association* **100**, 470–483.

[*Received December 5, 2005; accepted for publication February 7, 2006*]