Analysis of Case-Control Age-at-Onset Data Using a Modified Case-Cohort Method

Bin Nan[*, 1] and **Xihong Lin**[2]

[1] Department of Biostatistics, School of Public Health, University of Michigan,
1420 Washington Heights, Office: M4055 SPH II, Ann Arbor, MI 48109-2029, USA

[2] Department of Biostatistics, Harvard School of Public Health, 655 Huntington Avenue,
Building II Room 419, Boston, MA 02115, USA

Received 7 June 2007, revised 12 November 2007, accepted 18 December 2007

Summary

Case-control designs are widely used in rare disease studies. In a typical case-control study, data are collected from a sample of all available subjects who have experienced a disease (cases) and a subsample of subjects who have not experienced the disease (controls) in a study cohort. Cases are oversampled in case-control studies. Logistic regression is a common tool to estimate the relative risks of the disease with respect to a set of covariates. Very often in such a study, information of ages-at-onset of the disease for all cases and ages at survey of controls are known. Standard logistic regression analysis using age as a covariate is based on a dichotomous outcome and does not efficiently use such age-at-onset (time-to-event) information. We propose to analyze age-at-onset data using a modified case-cohort method by treating the control group as an approximation of a subcohort assuming rare events. We investigate the asymptotic bias of this approximation and show that the asymptotic bias of the proposed estimator is small when the disease rate is low. We evaluate the finite sample performance of the proposed method through a simulation study and illustrate the method using a breast cancer case-control data set.

Key words: Age-at-onset; Asymptotic bias; Bootstrap; Case-cohort; Case-control; Rare disease.

1 Introduction

Case-control designs are widely used as a cost-effective vehicle to study risk factors of a rare disease. In a typical case-control study, data are collected from all available subjects who have experienced a certain disease (cases) and a sub-sample of subjects who have not experienced the disease (controls) in a study cohort. Cases are oversampled in case-control studies. Since the outcome-based biased sampling nature can be ignored (see e.g. Prentice and Pyke, 1978), logistic regression is commonly used to estimate the associations of the presence/absence of the disease and a set of covariates measured by odds ratios as approximations of relative risks.

Age-at-onset of disease for all the cases and age at the survey for all the selected controls are often known in case-control studies where incident cases are recruited. A traditional analysis is to use age as a covariate in logistic regression. A model that handles time-to-event data would be more natural and appropriate, however, since it uses the information in the data more efficiently than simply dichotomizing subjects according to the presence/absence of the disease at a certain time point (see e.g. Cox and Oakes, 1984, Chapter 1), especially when censoring is present for controls on ages-at-onset and censoring times vary between subjects.

[*] Corresponding author: e-mail: bnan@umich.edu, Phone: 001 734 763 5538, Fax: 001 734 763 2215

If each case in a case-control study has one or more age matched controls whose ages are the same as the age at disease of the case, then the regression parameters of exposure variables in a Cox model can be estimated using the conditional likelihood method proposed by Prentice and Breslow (1978). Such a conditional likelihood method has recently been extended to family studies with correlated failure times, see Li, Yang, and Schwartz (1998), Hsu et al. (1999), and Shih and Chatterjee (2002), among others.

Many case-control studies are not age-matched, and yet the information of age-at-onset of disease is available. Our research is motivated by a breast cancer case-control study conducted at the University of Michigan (Beebe, 2002). The study consisted of 204 incident cases and 246 controls who were postmenopausal women. The major question of interest was how a woman's weight change was associated with the risk of breast cancer. The investigators collected ages at the breast cancer diagnosis for cases and ages at the survey for controls, but ages were not individually matched between cases and controls. For this type of unmatched (only for age) case-control studies, which is widely used in epidemiologic research, the conditional likelihood approach does not apply unless a very fine post hoc stratification on age and other continuous confounders can be reasonably applied, see Breslow and Day (1980) and Neuhänser and Becher (1997).

In this article, we propose to analyze such case-control age-at-onset data using a modified case-cohort method by treating the group of controls as an approximation of a subcohort under the assumption of a low population disease rate. The proposed method does not involve any post hoc stratification. In a case-cohort study (see e.g. Prentice, 1986; Self and Prentice, 1988), complete information is collected for all cases and all subjects in a subcohort that is a random sub-sample of the study cohort. The intuition behind treating case-control data as approximated case-cohort data is that the number of cases in the subcohort is close to zero for a rare disease study. We refer to Langholz and Goldstein (1996) for an overview of case-control, case-cohort and other risk set sampling methods.

We introduce the proposed method in Section 2. In Section 3, we perform an asymptotic bias analysis of the proposed method and show that the asymptotic bias of the relative risk estimator is very small when the disease rate is low, which is a common assumption underlying case-control designs. Numerical examples are given in Section 4. Simulations show that the proposed method works well for finite samples. We apply the method to the Michigan breast cancer case-control study, followed by discussions in Section 5. We use a nonparametric bootstrap method for variance estimation. As we point out in Section 5, our approach is ready to be extended to covariate matched case-control designs and family studies.

2 The Modified Case-Cohort Method

2.1 The case-control age-at-onset data

Consider a case-control study with n subjects. To follow the traditional notation used in case-cohort data analysis as described in Section 2.2, let $\Delta_i = 1$ if subject i is a case and 0 if a control, and Y_i be his/her age and Z_i be a vector of other covariates. Note that Y_i is age at disease diagnosis if subject i is a case and age at the survey if a control. Standard logistic regression has been commonly used for analyzing such case-control data using Δ_i as a binary outcome and age Y_i and covariates $\mathbf{Z}_i$ as independent variables in the light of the results of Prentice and Pyke (1978). Specifically, such a logistic model can be written as

$$\operatorname{logit}(p_i) = \beta_0 + \beta_1 Y_i + \beta_2' \mathbf{Z}_i, \tag{1}$$

where p_i is the probability of being a case given a subject is sampled into the case-control sample. The results of Prentice and Pyke (1978) show that parameters (β_1, β_2') are the odds ratios and can be estimated by fitting such a logistic regression to the case-control data, which are approximations of the population relative risks.

Since this traditional logistic regression uses the dichotomous disease status Δ_i as an outcome and does not fully use age-at-onset information of cases and censoring information of controls Y_i, instead, the age-at-onset variable (subject to censoring) Y_i is used as a covariate in (1). Given such information, it is more natural to analyze such data as survival data by treating age-at-disease-onset as a survival outcome and case-control status as a censoring indicator. A major difficulty of such an analysis is the presence of biased sampling with unknown selection probabilities in case-control data where cases are oversampled. To overcome this, we view such data as an approximation of data from a case-cohort study assuming the population disease rate is low and propose an analysis using a modified case-cohort method.

2.2 The estimation method of a case-cohort study

We first briefly review the estimating method of Self and Prentice (1988) for a case-cohort study, then show how it can be modified to analyze case-control age-at-onset data in the next subsection. Suppose a underlying study cohort consists of m independent subjects, where m is usually unknown in a case-control study. The disease status is known for every subject in the cohort. In a case-cohort study, complete information including covariates and event time subject to censoring is collected for all cases and all subjects in a random subsample of the study cohort, the so-called subcohort. Note that the cases in the subcohort are a subset of all the cases and hence the intersection of the subcohort and the set of cases may not be empty.

To be comparable to the case-control study described in the previous subsection, we assume there are n subjects in the case-cohort study. For subject i, let $Y_i \equiv \min(T_i, C_i)$ be the observed time where T_i is the failure time and C_i is the censoring time, $\Delta_i \equiv I(T_i \leq C_i)$ the failure indicator, and $\mathbf{Z}_i$ a vector of covariates. Assume the population follows the Cox model

$$\lambda(t \mid \mathbf{Z}) = \lambda_0(t) \exp(\theta \mathbf{Z}). \tag{2}$$

where $\lambda(\cdot)$ is the hazard function and $\lambda_0(\cdot)$ is the baseline hazard function of the failure time. Self and Prentice (1988) proposed to estimate the regression coefficients θ by solving the following estimating equation

$$\frac{1}{m}\sum_{i=1}^{m} \int \left\{ \mathbf{Z}_i - \frac{\sum_{j \in \mathcal{SC}} I(Y_j \geq t)\mathbf{Z}_j \exp(\boldsymbol{\theta}'\mathbf{Z}_j)}{\sum_{j \in \mathcal{SC}} I(Y_j \geq t) \exp(\boldsymbol{\theta}'\mathbf{Z}_j)} \right\} \mathrm{d}N_i(t) = 0\,, \tag{3}$$

where $\mathcal{SC}$ denotes the subcohort, $N_i(t) \equiv \Delta_i I(Y_i \leq t)$ is the failure counting process for subject i, and m is the total number of subjects in the underlying study cohort. A detailed discussion of m is given in the next subsection. Self and Prentice (1988) proved that the estimator obtained from Eq. (3) is consistent and asymptotically normal. If $\mathcal{SC}$ is the entire cohort, Eq. (3) becomes the partial likelihood estimating equation for cohort data.

The subcohort $\mathcal{SC}$ in Eq. (3) is a simple random sample of the study cohort. One can easily see that the ratio inside the integral of Eq. (3) is an estimator of the following quantity

$$\frac{E_0\{I(Y \geq t)\mathbf{Z} \exp(\boldsymbol{\theta}'\mathbf{Z})\}}{E_0\{I(Y \geq t) \exp(\boldsymbol{\theta}'\mathbf{Z})\}}\,, \tag{4}$$

where E_0 denotes the expectation taken under the Cox model (2) at the true parameter value. The uniform consistency of the ratio in Eqs. (3) to (4) can be shown by standard empirical process arguments under the assumptions in the Appendix (see e.g. van der Vaart and Wellner, 1996).

2.3 The modified case-cohort method for the case-control age-at-onset data

If we let P be the joint distribution of a single observation from the underlying population and P_1 and P_0 be the conditional distributions of the observation given $\Delta = 1$ and $\Delta = 0$, respectively, then a case-control study consists of two independent random samples from P_1 and P_0, respectively, whereas a case-cohort study consists of two random samples (can be overlapped) from P_1 and P, respectively.

In a case-cohort study, the sample from P (subcohort) is used to estimate (4), which is given in Eq. (3).

For a rare disease, the probability of observing $\Delta = 1$ is very small, hence the number of cases would be very small compared to the number of controls in a subcohort if it were available. We thus can treat the group of all available controls in a case-control study as the controls arising from an underlying subcohort $\mathcal{SC}$. Under the rare disease assumption, we can use controls to approximately estimate (4). In other words, we use P_0 to approximate P for estimating (4). This view allows us to treat case-control data as approximated case-cohort data and estimate the regression coefficients $\boldsymbol{\theta}$ by approximating the ratio in the case-cohort estimating equation (3) using only controls.

Specifically, assuming the underlying population that the case-control data are generated from follows the Cox model (2), based on Eq. (3) we propose the following estimating equation for case-control data

$$\frac{1}{m}\sum_{i=1}^{m}\int\left\{\mathbf{Z}_i - \frac{\sum_{j\in\mathcal{SC}}(1-\Delta_j)\,I(Y_j \geq t)\,\mathbf{Z}_j \exp{(\boldsymbol{\theta}'\mathbf{Z}_j)}}{\sum_{j\in\mathcal{SC}}(1-\Delta_j)\,I(Y_j \geq t)\exp{(\boldsymbol{\theta}'\mathbf{Z}_j)}}\right\} \mathrm{d}N_i(t) = 0\,, \tag{5}$$

where m indicates the underlying cohort size, $\mathcal{SC}$ denotes the "underlying" subcohort, Y_i is the age of subject i, $\Delta_i = 1$ if subject i is a case and 0 if a control. This notation is consistent with that in Section 2.1. Note that (5) differs from the case-cohort estimating equation (3) by the additional factor $(1 - \Delta_i)$ in both the numerator and the denominator of the second term in the integrand. Since $1 - \Delta_j = 1$ for controls and 0 for cases, one can easily see that the numerator and the denominator of the ratio on the left hand side of (5) only sum over controls and are hence fully determined by the observed case-control data.

The underlying cohort size m is often unknown for case-control studies. Equation (5), however, is completely determined since only the summands with $\mathrm{d}N_i(t) = 1$ contribute to the left hand side of (5) and all these subjects are cases and hence fully observed. We keep m in the formula just for notational convenience and ease of describing asymptotic properties. For implementation purpose, denoting the set of cases by $\mathcal{C}$ in the case-control sample, and the set of *controls* who are at risk for case i by $\mathcal{R}_i$, Eq. (5) can be simplified to

$$\sum_{i\in C}\left\{\mathbf{Z}_i - \frac{\sum_{j\in\mathcal{R}_i}\mathbf{Z}_j \exp{(\boldsymbol{\theta}'\mathbf{Z}_j)}}{\sum_{j\in\mathcal{R}_i}\exp{(\boldsymbol{\theta}'\mathbf{Z}_j)}}\right\} = 0\,.$$

In a case-cohort study, the proportion of the subcohort size in the whole cohort is needed for estimating the variance of the estimator of θ obtained from solving Eq. (3), see e.g. Self and Prentice (1988). Such information, however, is usually unknown for a case-control age-at-onset data. We propose to use nonparametric bootstrap to estimate the variance of $\hat{\boldsymbol{\theta}}$ obtained by solving Eq. (5). We generate bootstrap samples by resampling from cases and controls separately.

Note that estimating Eq. (5) can be deduced from Eq. (3.5) of Chen and Lo (1999) by taking the population case percentage to be zero, which in fact can be traced back to the weighted estimating method of Kalbfleisch and Lawless (1988). The asymptotic result of Chen and Lo (1999), however, does not apply because (i) their corresponding asymptotic variance given in (3.6c) would contain a division of $1/0$ if a zero case percentage were plugged in, and (ii) their asymptotic variance in (3.6c) seems to be incorrect since it yields an over 100% asymptotic relative efficiency (see the last row of column 2 in their Table 1). The estimator obtained from their Eq. (3.5) should not be super-efficient.

3 The Asymptotic Bias of the Modified Case-Cohort Estimator

It is of significant practical interest to investigate the performance of the case-cohort approximation for case-control age-at-onset data. Here we study the asymptotic bias of the estimator obtained from

Eq. (5). We consider a one-dimensional covariate here for notational simplicity. Denote by $\hat{\theta}$ the solution of Eq. (5), and by θ_0 the true value of θ. Let

$$\eta(t;\theta) = \frac{E_0\{(1-\Delta)\, I(Y \geq t)\, Z \exp(\theta Z)\}}{E_0\{(1-\Delta)\, I(Y \geq t) \exp(\theta Z)\}} \,. \tag{6}$$

We show in the Appendix that the left hand side of the Eq. (5) is asymptotically equivalent to

$$\frac{1}{m} \sum_{i=1}^{m} \int \{Z_i - \eta(t;\theta)\}\, \mathrm{d}N_i(t) = \frac{1}{m} \sum_{i=1}^{m} \{Z_i - \eta(Y_i;\theta)\}\, \Delta_i \,, \tag{7}$$

which converges to

$$\psi(\theta) = E_0[\{Z - \eta(Y;\theta)\}\, \Delta] \tag{8}$$

in probability uniformly, where the expectation $E_0(\cdot)$ is taken under the true model that the data (Y, Δ, Z) follow, i.e. the Cox model (2). The asymptotic limit θ_* of $\hat{\theta}$ hence solves $\psi(\theta_*) = 0$.

Specifically, under the Cox model (2), let $F_0(\cdot \mid z)$ and $G_0(\cdot \mid z)$ denote the conditional distribution functions of T and C given $Z = z$, respectively, and $f_0(\cdot \mid z)$ and $g_0(\cdot \mid z)$ the corresponding density functions. Let $h_0(\cdot)$ be the density function of Z. Assuming T and C are independent conditional on Z, i.e., independent censoring, the joint density function of (Y, Δ, Z) is

$$p_0(y,\delta,z;\theta_0,\lambda_{00}) = [\{1 - G_0\,(y \mid z)\} f_0(y \mid z)]^{\delta}\, [\{1 - F_0(y \mid z)\}\, g_0(y \mid z)]^{1-\delta}\, h_0(z),$$

where θ_0 denotes the true value of θ, $\lambda_{00}(\cdot)$ denotes the true baseline hazard, $\Lambda_{00}(t)$ is the true cumulative baseline hazard and $F_0(t \mid z) = \exp\{-\Lambda_{00}(t) \exp(\theta_0' z)\}$. Hence Eq. (8) can be written as

$$\begin{aligned}\psi(\theta;\theta_0,\lambda_{00}) &= \sum_{\delta=0}^{1} \int\!\!\int \{z - \eta(y;\theta)\}\, \delta p_0(y,\delta,z;\theta_0,\lambda_{00})\, \mathrm{d}y\, dz \\ &= \int\!\!\int \{z - \eta(y;\theta)\}\{1 - G_0(y \mid z)\} f_0(y \mid z)\, h_0(z)\, \mathrm{d}y\, dz.\end{aligned}$$

It follows that the asymptotic limit θ_* solves $\psi(\theta_*;\theta_0,\lambda_{00}) = 0$ and hence θ_* is a function of the true values (θ_0, λ_{00}). Numerical integration can be used to calculate $\eta(\cdot)$ and $\psi(\cdot)$ and the Newton-Raphson algorithm can be used to numerically solve the equation.

Now we calculate the asymptotic bias of $\hat{\theta}$ numerically. We assume the baseline failure time T has an exponential distribution with constant hazard rate λ_{00}. Thus $\Lambda_{00}(t) = \lambda_{00} t$. Assume the censoring time C follows a uniform distribution in $(0, c_0)$ and the follow-up ends at a fixed time τ with $\tau < c_0$, and the covariate is binary, i.e., $Z \in \{0, 1\}$. We choose $\tau = 1$, $c_0 = 2$, and $\Pr(Z = 1) = 0.5$. We choose different values of λ_{00} corresponding to different values of the disease rate d_{00} of the $Z = 0$ group at time τ in absence of censoring. In other words, $\lambda_{00} = -\log(1 - d_{00})$. We calculate the asymptotic relative bias assuming the true value of θ_0 is $\pm\log 2 \approx \pm 0.693$, 0, $\pm\log 3 \approx \pm 1.099$, which correspond to hazard ratios 2 and 1/2, 1, and 3 and 1/3. The asymptotic relative bias is defined as $(\theta_* - \theta_0)/\mid \theta_0 \mid$. When $\theta_0 = 0$, the relative bias is calculated as $\theta_* - \theta_0$.

Under the assumed model, the function $\eta(t;\theta)$ can be calculated analytically using the above joint density function $p_0(\cdot)$. We then calculate the function $\psi(\theta)$ and its derivative with respect to θ using numeric differentiation, and use the Newton-Raphson algorithm to search for the root of $\psi(\theta_*) = 0$. We find the derivative of $\psi(\theta)$ is always negative, and thus $\psi(\theta)$ decreases monotonically and the root of $\psi(\theta_*) = 0$ is unique. Figure 1 presents the asymptotic relative bias as a function of the baseline disease rate for the five values of θ_0. We vary the baseline disease rate from close to 0% to 5%. The 5% baseline disease rate corresponds to the population disease rate 9.6%, 7.4%, 5%, 3.8%, and 3.3% for $\theta_0 = \log 3$, $\log 2$, 0, $-\log 2$, and $-\log 3$, respectively. The results in Figure 1 show that the asymptotic bias of $\hat{\theta}$ is anti-conservative, and the relative bias is very small and is up to 4% when the population disease rate is less than 10%.

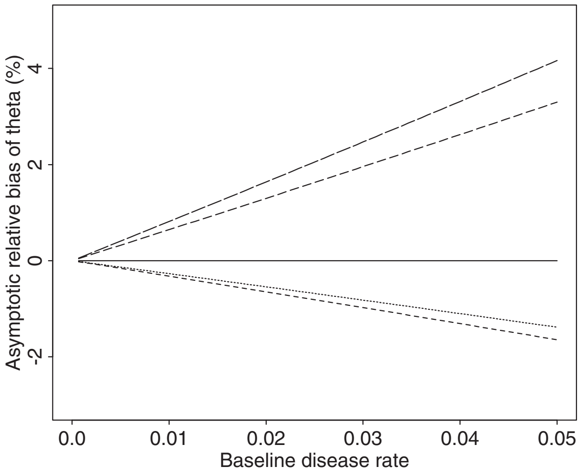

Figure 1 Asymptotic bias of θ as a function of the baseline disease rate at different values of θ_0: - - - - - $\theta_0 = -\ln(3)$; . . . . . . $\theta_0 = -\ln(2)$; ——— $\theta_0 = 0$; – – – $\theta_0 = \ln(2)$; — — — $\theta_0 = \ln(3)$.

4 A Simulation Study

We have conducted a simulation study to investigate the finite sample performance of the proposed method. Data are generated from the same distribution as that used in the above theoretical asymptotic bias calculation. For each pair of the baseline disease rate d_{00} and θ_0, we randomly generate a large cohort and randomly select 100 failures (cases) prior to time $\tau = 1$ and 100 controls. The baseline disease rate is set to be 0.5% and 2.5%. They correspond to the population disease rate between 0.3% to 4.9% when θ varies between $-\log(3)$ and $\log(3)$. For each parameter configuration, we simulate 1000 data sets, and analyze each data set using the proposed modified case-cohort method by solving Eq. (5). The numerical implementation follows the method proposed by Therneau and Li (1999). For comparison purpose, we have also performed a standard logistic regression using the case/control status as a binary outcome and the binary indicator Z_i and time (age) Y_i as covariates. We use 100 bootstrap runs to calculate the bootstrap variance of $\hat{\theta}$ under the modified case-cohort analysis, where cases and controls are re-sampled with replacement separately.

Table 1 reports the empirical bias, the empirical standard error, and the average of the bootstrap standard errors of the modified case-cohort analysis and the average of the model-based standard errors of the logistic regression. The theoretical asymptotic bias for the modified case-cohort analysis is also reported. The results in Table 1 suggest that the proposed modified case-cohort method performs well for analyzing case-control age-at-onset data. The empirical bias is very small and agrees with the asymptotic bias reasonably well. The bootstrap standard errors agree with their empirical counterparts. The empirical biases and standard errors of the modified case-cohort survival analysis are smaller than those from standard case-control logistic analysis, especially when the disease rate is lower, indicating a superior performance of the proposed modified case-cohort method for rare diseases.

5 Analysis of the Breast Cancer Case-Control Data

We have applied the proposed method to analyzing data from the Michigan breast cancer case-control study (Beebe, 2002). The study consisted of 204 cases and 246 controls who were postmenopausal women. Cases were identified from patients' records at the University of Michigan Breast Cancer

Table 1 Biases and standard errors based on 1000 simulations with 100 cases and 100 controls in each data set using the modified case-cohort (MCC) analysis Eq. (5) and the standard logistic regression (SLR).

θ_0	Population desease rate	*Baseline disease rate = 0.5%*				Bstrp. SE Model SE
		Method	Asym. bias	Empr. bias	Empr. SE	
$-\log(3)$	0.3%	MCC	−0.001	−0.026	0.317	0.328
		SLR	–	−0.047	0.344	0.351
$-\log(2)$	0.4%	MCC	−0.001	−0.002	0.307	0.311
		SLR	–	−0.017	0.335	0.332
0	0.5%	MCC	0	0.006	0.301	0.302
		SLR	–	0.009	0.331	0.322
$\log(2)$	0.7%	MCC	0.002	0.008	0.313	0.312
		SLR	–	0.009	0.345	0.333
$\log(3)$	1%	MCC	0.004	0.012	0.323	0.327
		SLR	–	0.035	0.359	0.350

θ_0	Population disease rate	*Baseline disease rate = 2.5%*				Bstrp. SE Model SE
		Method	Asym. bias	Empr. bias	Empr. SE	
$-\log(3)$	1.7%	MCC	−0.008	−0.043	0.323	0.328
		SLR	–	−0.044	0.365	0.350
$-\log(2)$	1.9%	MCC	−0.006	−0.001	0.318	0.311
		SLR	–	−0.006	0.351	0.332
0	2.5%	MCC	0	−0.009	0.303	0.303
		SLR	–	−0.012	0.329	0.321
$\log(2)$	3.7%	MCC	0.011	0.017	0.308	0.313
		SLR	–	0.012	0.340	0.331
$\log(3)$	4.9%	MCC	0.023	0.026	0.328	0.327
		SLR	–	0.026	0.359	0.348

Center, who were diagnosed with primary breast cancer between January 1, 1996 and December 31, 1999. The major question of interest was how a woman's weight change was associated with the risk of breast cancer. The investigators collected ages at the breast cancer diagnosis for cases and ages at the survey for controls. All the participants were white. The ages of the participants ranged between 50 and 70 for either cases or controls, with medians equal to 57 for cases and 60 for controls. The covariate of main interest was the change of body mass index between age 20 to 50. Each woman was asked to report her weights at age 20 and 50, marital status, smoking status, family history of breast cancer and birth of a child. Body mass indexes (BMI) at age 20 and 50 were calculated (weight/height^2, kg/m^2), and the change of BMI between age 50 and 20 was calculated. Among those 450 cases and controls, 21 women with missing information on either weight change or height were excluded from the analysis. One additional extremely short woman (40 inches high, which was likely to be a coding error) was also excluded.

We analyzed the data using the modified case-cohort method under the Cox model by solving (5), where ages of cases were used as event times and ages of controls were used as censoring times. The breast cancer prevalence percent of January 1, 2003 of SEER 11 population diagnosed in the previous

Table 2 Analyses of the Michigan Breast Cancer Case-Control Data.

Variable	Modified case-cohort		Logistic regression		Conditional likelihood	
	Coefficient	SE	Coefficient	SE	Coefficient	SE
Change in BMI	0.03	0.09	−0.16	0.07	−0.12	0.07
BMI at Age 20	−0.05	0.04	−0.08	0.03	−0.07	0.03
Marital Status	−0.67	0.39	−0.95	0.28	−0.91	0.28
Smoking Status	0.90	0.32	0.31	0.22	0.36	0.22
Family history	1.26	0.39	1.11	0.27	1.01	0.27
Child Birth	−1.07	0.42	−0.41	0.33	−0.32	0.34

10 years for white women aged between 50 and 70 is less than 3% (see Ries et al., 2006). The prevalence rate for the University of Michigan breast cancer data set should be even smaller because of shorter cancer diagnostic period. Hence our rare disease assumption is appropriate. Our asymptotic bias analysis (Section 3) and simulation results (Section 4) show that the proposed modified case-cohort method is expected to work well for such a rare disease. We calculated standard errors using 1000 bootstraps. The covariates included in the model were change of BMI, baseline BMI, marital status (yes/no), current smoking status (yes/no), family history (yes/no) and birth of a child (yes/no). The results are presented in Table 2. The change in BMI did not show a significant association with the risk of breast cancer. Smoking and family history of breast cancer significantly increased the risk of breast cancer. While marriage only marginally decreased the risk (with p-value < 0.1), child birth significantly decreased the risk of breast cancer. No significant relationship was found between baseline BMI, BMI change and breast cancer.

We also did a stratified analysis with a 10-year age window as a stratum to take into account the possible birth cohort effect, and found similar results as the non-stratified results given in Table 2. Thus we only report the non-stratified analysis in this article.

In Table 2 we also included results of a logistic regression adjusted for age and a conditional likelihood estimation with post hoc stratification on age. We have tried different functional forms of age in the logistic regression including cubic spline smoothing and observed similar results. Hence we only reported the result adjusted by linear age effect. The post hoc stratification is based on one-year intervals. The logistic regression and the post hoc stratification yield similar results, but their effects of BMI change is different to that obtained from the proposed method. Effects of BMI at age 20, smoking status, and child birth have also switched significance status. We do not have a definitive explanation on the discrepancy in addition to the methodological differences. There is always a chance that some important confounders were missed in the study design.

6 Discussions

The modified case-cohort survival analysis provides an attractive procedure for analyzing case-control age-at-onset data when the disease rate is low. Unlike the standard logistic regression used for case-control data, this analysis naturally uses age at disease onset as a survival outcome and is easy to implement. Our simulation results show that the proposed method outperforms standard case-control logistic regression when the disease rate is low. In view of the lack of the sampling fraction of controls in case-control data, we propose to estimate standard errors using bootstrap. It is of future research interest to develop an alternative analytic standard error estimator. The proposed approach can handle time-dependent covariates without any added difficulty.

We restrict in this paper on classical case-control data where ages are not matched. The proposed method can also be generalized to exposure covariate matched case-control data and family case-control data. For the exposure covariate matched case-control data, each group of cases and controls with matched covariates consists a stratum. Then a stratified Cox regression can be implemented together with the method of Therneau and Li (1999), which assumes each stratum has its own baseline hazard function. For the correlated family data, the method of Cai and Prentice (1995) may apply to improve estimation efficiency for hazard ratio parameters. Note that the cumulative baseline hazard function is not estimable from case-control data where age is not matched because the sampling fraction of controls is unknown.

Appendix: Proof of the Asymptotic Limit of Eq. (5)

Again we consider one-dimensional covariate here for notational simplicity. The Multidimensional case can be handled similarly. Let

$$\hat{\eta}(t;\theta) = \frac{\sum_{j\in\mathcal{SC}} (1-\Delta_j)\, I(Y_j \geq t)\, Z_j \exp(\theta Z_j)}{\sum_{j\in\mathcal{SC}} (1-\Delta_j)\, I(Y_j \geq t) \exp(\theta Z_j)}\,. \tag{9}$$

Suppose the parameter space θ is compact, and the study ends at a fixed time τ with $\Pr(C \geq \tau) = \Pr(C = \tau) > 0$ while $\Pr(T > \tau) > 0$. Then all functions $1 - \Delta$, $I(Y \geq t)$, Z and $\exp(\theta Z)$ are well-behaved and belong to Donsker classes. Hence both summands in the numerator and the denominator in $\hat{\eta}(t,\theta)$ are in Donsker classes and thus belong to Glivenko Cantelli classes (van der Vaart and Wellner, 1996). It follows that $\hat{\eta}(t;\theta)$ converges to $\eta(t;\theta)$ in probability uniformly in $[0,\tau] \times \theta$ when the size of control group approaches to infinity, i.e., as $m \to \infty$, given that the denominator is bounded away from zero in probability. Then the left hand side of equation (5) is asymptotically equivalent to (7) uniformly in θ since

$$\left|\frac{1}{m}\sum_{i=1}^{m} \int \{\hat{\eta}(t;\theta) - \eta(t;\theta)\}\, \mathrm{d}N_i(t)\right| \leq \sup_{t,\theta} \mid \hat{\eta}(t;\theta) - \eta(t;\theta) \mid \cdot \frac{1}{m}\sum_{i=1}^{m} \Delta_i \to 0$$

in probability by the uniform convergence of $\hat{\eta}$. By the permanence of the Donsker property for convex hulls of van der Vaart and Wellner (1996), it can be shown that $\{\eta(t;\theta)\}$ is also Donsker and thus a Glivenko-Cantelli class given that the denominator is bounded away from zero. Hence the right hand side of (7) converges to $\psi(\theta)$ in equation (8) in probability uniformly in Θ.

Acknowledgements The authors would like to thank Editor Edgar Brunner, the Associate Editor and three referees for their constructive comments, and to Jennifer Beebe for providing the breast cancer data. The research was supported in part by U.S. National Cancer Institute grant R01 CA76404 and U.S. NSF grant DMS-07-06700.

References

Beebe, J. (2002). Body Mass at different periods, adult weight gain and risk of breast cancer and endometrial cancer. Ph.D. dissertation, University of Michigan.

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research Vol. I: The analysis of Case-Control Studies*. IARC Scientific Publications no. 32, Lyon.

Ca, J. and Prentice, R. L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika* **82**, 151–164.

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

Cox, D. R. and Oakes, D. (1984). *Analysis of Survival Data*. Chapman and Hall/CRC, Boca Raton.

Hsu, L., Prentice, R. L., Zhao, L. P., and Fan, J. J. (1999). On dependence estimation using correlated failure time data from case-control family studies. *Biometrika* **86**, 743–753.

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

Langholz, B. and Goldstein, L. (1996). Risk set sampling in epidemiologic cohort studies. *Statistical Science* **11**, 35–53.

Li, H., Yang, P., and Schwartz, A. G. (1998). Analysis of age of onset data from case-control family studies. *Biometrics* **54**, 1030–1039.

Neuhäuser, M. and Becher, H. (1997). Improved odds ratio estimation by post hoc stratification of case-control data. *Statistics in Medicine* **16**, 993–1004.

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

Prentice, R. L. and Breslow, N. E. (1978). Retrospective studies and failure time models. *Biometrika* **65**, 153–158.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika* **66**, 403–411.

Ries, L. A. G., Harkins, D., Krapcho, M., Mariotto, A., Miller, B. A., Feuer, E. J., Clegg, L., Eisner, M. P., Horner, M. J., Howlader, N., Hayat, M., Hankey, B. F., Edwards, B. K. (eds). (2006). *SEER Cancer Statistics Review, 1975–2003*. National Cancer Institute. Bethesda, MD, http://seer.cancer.gov/csr/1975_2003/, based on November 2005 SEER data submission, posted to the SEER web site.

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

Shih, J. H. and Chatterjee, N. (2002). Analysis of survival data from case-control family studies. *Biometrics* **58**, 502–509.

Therneau, T. M. and Li, H. (1999). Computing the Cox model for case cohort designs. *Lifetime Data Analysis* **5**, 99–112.

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