Published in final edited form as:

Biometrics. 2015 December ; 71(4): 960–968. doi:10.1111/biom.12342.

Retrospective Likelihood Based Methods for Analyzing Case-Cohort Genetic Association Studies

Yuanyuan Shen¹, Tianxi Cai¹, Yu Chen², Ying Yang³, and Jinbo Chen⁴

¹Department of Biostatistics, Harvard University, Boston, MA 02115

²Department of Population Health, New York University School of Medicine New York, NY 10016

³Department of Mathematical Sciences, Tsinghua University, Beijing 100084, P. R. China

⁴Department of Biostatistics and Epidemiology, University of Pennsylvania School of Medicine, Philadelphia, PA 19104

Summary

The case cohort (CCH) design is a cost effective design for assessing genetic susceptibility with time-to-event data especially when the event rate is low. In this work, we propose a powerful pseudo score test for assessing the association between a single nucleotide polymorphism (SNP) and the event time under the CCH design. The pseudo score is derived from a pseudo likelihood which is an estimated retrospective likelihood that treats the SNP genotype as the dependent variable and time-to-event outcome and other covariates as independent variables. It exploits the fact that the genetic variable is often distributed independent of covariates or only related to a low-dimensional subset. Estimates of hazard ratio parameters for association can be obtained by maximizing the pseudo likelihood. A unique advantage of our method is that it allows the censoring distribution to depend on covariates that are only measured for the CCH sample while not requiring the knowledge of follow up or covariate information on subjects not selected into the CCH sample. In addition to these flexibilities, the proposed method has high relative efficiency compared with commonly used alternative approaches. We study large sample properties of this method and assess its finite sample performance using both simulated and real data examples.

Keywords

Case-cohort design; Cox proportional hazards model; Genetic association; Inverse probability weighting; Pseudo-likelihood; Polytomous regression

1. Introduction

The CCH design is a cost-effective two-phase sampling strategy that has been widely used for assessing the association between costly exposures and time-to-event outcomes (Prentice, 1986). Under this design, expensive exposures are only measured on subjects who

*yyang@math.tsinghua.edu.cn.

5. Supplementary Materials

Web Appendices referenced in Section 2.1 and 2.2 and the source code for computation are available with this paper at the *Biometrics* website on Wiley Online Library.

experienced the interest event (cases) and a random subset of the study cohort. Recently, this design has been adopted to study genetic susceptibility in relation to human disease phenotypes (e.g., Christensen et al, 2008; Silander et al, 2008), where data for non-genetic risk factors were collected for the full cohort. But genotype data, either for a set of candidate genes or for the whole genome, was obtained only for the CCH subsample. With outcome dependent sampling appropriately accounted for, the CCH data permits unbiased assessment of gene-phenotype associations. Genotype data in the random subsample permits population genetic data analysis, such as estimation of allele and haplotype frequencies and assessment of Hardy-Weinberg proportions.

A wide range of statistical methods have been developed for estimating hazard ratio parameters with CCH data, which can be applied to estimate and test SNP effects. These methods can largely be classified into three classes: (I) weighted likelihood or weighted estimating equation methods (Kalbfleisch and Lawless, 1988; Chen and Lo, 1999; Borgan et al. estimator II, 2000; Chen, 2001; Kulich and Lin, 2004; Nan, Emond and Wellner, 2004; Qi, Wang, and Prentice, 2006), (II) pseudo-likelihood methods (Prentice, 1986; Self and Prentice, 1988; Barlow, 1994; Borgan et al. estimators I and III, 2000), and (III) nonparametric maximum likelihood estimation (NPMLE) methods (Chen and Little, 1999; Scheike et al., 2004). These methods are expected to have similar statistical efficiencies for assessing genetic associations if they rely on incorporating auxiliary variables that are correlated with the interest exposure to achieve improved efficiency. This is because correlations between genotypes and auxiliary non-genetic variables are often weak. For example, when events are rare, the hazard ratio parameters are small or moderate, and auxiliary non-genetic variables are only weakly correlated with genotypes, limited efficiency improvement can be achieved by incorporating auxiliary information in inverse probability weighted (IPW) estimators (Samuelsen et al., 2007). For the NPMLE methods, the censoring is allowed to be dependent on fully observed variables, but not on variables collected only for the CCH sample. They also require the estimation of the joint conditional distribution of expensive markers collected only on the CCH sample given the covariates measured on the entire cohort, which could be challenging when continuous variables are involved.

We propose both testing and estimation procedures for assessing the association between a SNP, G , and a event time T in the presence of additional covariates W under a CCH design. We focus on the setting where no information on the subjects outside of the CCH subcohort is required beyond the size of the study cohort. This is desirable when extracting information on the entire large cohort is difficult such as when ascertaining the clinical variables W is costly or when the CCH sample is assembled from a consortium. Existing methods that require the knowledge of follow up and W information on the entire cohort, such as the NPMLE, are not applicable to such settings. Our proposed procedures are based on a retrospective likelihood derived from the Cox proportional hazards model for T given G and W , with G treated as the dependent variable. Our method allows the censoring distribution to depend on W and only requires information from the CCH subcohort in addition to the sampling fractions. Therefore, our method requires less stringent assumption compared to the NPMLE.

The remaining of this article is organized as follows. In Section 2, we describe a pseudo score test statistic and a pseudo-likelihood method for estimating hazard ratio parameters and provide large sample properties. In Section 3, we conduct extensive simulation studies and analyze a real data example to evaluate finite sample properties of our method relative to existing ones. We make conclusion remarks in Section 4.

2. The Proposed Methods

Let T denote the event time for the phenotype of interest, G denote the SNP genotype which takes values $0, 1, \dots, g_0$, and W denote the covariates. Due to censoring, T is only observable up to $X = \min(T, C)$ and $\delta = I(T \leq C)$, where C is the censoring time. Although SNP genotypes typically take values $0, 1$ or 2 , here we allow g_0 to be any integer to incorporate other scenarios such as dominant or recessive models. We assume that the censoring time C does not depend on G but can depend on covariates W , where W can include both discrete and continuous components. We assume that the underlying full cohort data

$\mathcal{D}_F = \{(X_i, \Delta_i, G_i, W_i'), i = 1, \dots, N\}$ are independent and identically distributed random vectors. Due to the CCH design, (X_i, G_i, W_i') is only observed if $\xi_i = 1$, where ξ_i indicates whether the i th subject is sampled into the CCH subset. For simplicity, we assume that Bernoulli sampling is employed with $\pi_i \equiv P(\xi_i = 1 | \mathcal{D}_F) = \frac{n_i}{n_0 + (1 - \frac{n_i}{n_0})N_0}$, where n_0 is a pre-specified number and N_0 is the total number of controls. Our method only requires information on subjects in $\Omega_{\text{CCH}} = \{i : \xi_i = 1\}$ for inference. Suppose that the main interest is to assess the association between G and T adjusting for covariates W . We adopt a Cox proportional hazards model to describe this association:

$$\lambda(t|W, G) = \lambda_0(t) \exp \left\{ \beta'_W W + \sum_{g=1}^{g_0} \beta_g I(G=g) \right\} = \lambda_0(t) \exp(\beta' Z), \quad (1)$$

where $\lambda(t|W, G)$ is the conditional hazard function of T given W and G , $\lambda_0(t)$ is an unspecified baseline hazard function, $\beta = (\beta'_W, \beta'_G)'$ are the hazard ratio parameters, $\beta_G = (\beta_1, \dots, \beta_{g_0})'$ represents the genetic effect, $Z = (W', I(G=1), \dots, I(G=g_0))'$. The gene-phenotype association is assessed by testing the null hypothesis

$$H_0: \beta_G = 0.$$

An analogy between CCH and case-control study designs can help motivate our proposed methods. Both designs select subjects based on the outcome variable. For the case-control design, standard statistical inference of odds ratio parameters is based on the retrospective likelihood (Prentice and Pyke, 1979) which is semiparametric efficient (Breslow et al., 2000). Suppose that counts of the total numbers of eligible cases and controls in the study population (from which cases and controls were ascertained) are available, so that the probability of sampling a case or control can be estimated. Then a weighted logistic regression analysis could also be performed, which may have lower statistical efficiency than the retrospective likelihood method (e.g., Breslow and Chatterjee, 1999). On the other

hand, with a sufficient number of controls, a case-control design coupled with the retrospective likelihood analysis could have very high statistical efficiency compared with the full cohort design. This motivated us to derive $p(G|X, W)$ for the CCH design, the probability of G conditioned on X and covariates W , and make inference on β_G based on the corresponding retrospective likelihood.

Let $\lambda_c(T_c|W)$ denote the hazard function for C given W , $\mathcal{D} = (X, W)'$,

$\eta = (\beta'_G, \alpha')'$, $\theta(t) = \{\beta'_W, \Lambda_0(t)\}$, and $\Lambda_0(X) = \int_0^X \lambda_0(t) dt$. Under (1) and the assumption that C is independent of G given W , it is not difficult to see that

$$\begin{aligned} \log \frac{P(G=g|\Delta, X, W)}{P(G=0|\Delta, X, W)} &= \log \frac{p(\Delta, X|G=g, W)}{p(\Delta, X|G=0, W)} + \log \frac{P(G=g|W)}{P(G=0|W)} \\ &= \log \frac{\lambda_0^\Delta(X) e^{\Delta(\beta_g + \beta'_W W)} e^{-\Lambda_0(X)} e^{\beta_g + \beta'_W W} \lambda_c^{1-\Delta}(X|W) e^{-\Lambda_c(X|W)}}{\lambda_0^\Delta(X) e^{\Delta\beta_W W} e^{-\Lambda_0(X)} e^{\beta'_W W} \lambda_c^{1-\Delta}(X|W) e^{-\Lambda_c(X|W)}} + \mathcal{R}_g(\alpha, W) \quad (2) \\ &= \Delta\beta_g - \Lambda_0(X) e^{\beta'_W W} (e^{\beta_g} - 1) + \mathcal{R}_g(\alpha, W) \equiv r_{\mathcal{D}}\{g|\eta, \theta(t)\}, \end{aligned}$$

where $P(G = g|W)$ is modeled through

$$\mathcal{R}_g(\alpha, W) = \log \frac{P(G=g|W)}{P(G=0|W)}, \quad g=1, \dots, g_0, \quad (3)$$

with α being the unknown model parameters. Here, additional modeling is required for $P(G = g|W)$ as in the NPMLE. In the absence of W , α simply is a vector of g_0 intercepts for $G = 1, \dots, g_0$. On the other hand, G may be correlated only with a subset of W , denoted by W_- . For example, W_- may include only indicator variables for ethnicity groups. With discrete W_- , α can be set as stratum-specific intercept parameters, where the stratum is defined by levels of W_- . When W_- has continuous components, a polytomous regression model may be adopted for $\mathcal{R}_g(\alpha, W) = \mathcal{R}_g(\alpha, W_-)$.

Based on (2), the log retrospective likelihood for the CCH data can be written as

$$\ell\{\eta, \theta(t)\} = \sum_{i=1}^{N_c} \left[r_{\mathcal{D}_i}\{G_i|\eta, \theta(t)\} - \log \left\{ 1 + \sum_{g=1}^{g_0} e^{r_{\mathcal{D}_i}\{g|\eta, \theta(t)\}} \right\} \right]. \quad (4)$$

The CCH sampling is naturally accounted for in the retrospective likelihood by conditioning on $(X, \cdot)$. Because $\theta(t)$ is unknown, the log-likelihood $\ell\{\eta, \theta(t)\}$ cannot be used directly to obtain an estimate of η . But likelihood (4) cannot be used to estimate $\theta(t)$ because $(X, \cdot)$ and W are conditioned out. We use the Chen and Lo (1999) IPW method and obtain

$\hat{\beta}^\pi = (\hat{\beta}_W^\pi, \hat{\beta}_1^\pi, \hat{\beta}_2^\pi)'$ as a weighted partial likelihood estimator under (1) and $\hat{\Lambda}_0^\pi(t)$ as a weighted Breslow's estimator,

$$\hat{\Lambda}_0^\pi(t) = \sum_{j \in \Omega_{\text{CCH}}} \frac{I(X_j \leq t) \Delta_j}{\sum_{l \in \Omega_{\text{CCH}}} \hat{\pi}_l^{-1} I(X_l \geq t) e^{Z' \hat{\beta}^\pi}}.$$

Although our procedure can be carried out with other estimators for $\theta(t)$, we focus on the

IPW estimator $\hat{\theta}^\pi(t) = \{\hat{\beta}_w^\pi, \hat{\Lambda}_0^\pi(t)\}$ for all our theoretical and numerical analysis.

Below, we propose a pseudo score statistic for testing H_0 and a pseudo likelihood method for estimating β_G . For the ease of presentation, we estimate $\theta(t)$ as $\hat{\theta}^\pi(t)$ under the full model (1) for both testing and estimation but note that one may also estimate $\theta(t)$ under the null model and the proposed methods can be applied accordingly.

2.1 A pseudo score statistic for testing the null hypothesis $\beta_G = 0$

Under $H_0 : \beta_G = 0$, the log-likelihood score function for β_g , $g = 1, \dots, g_0$, based on (4) can be written as

$$U_{\beta_g}\{\alpha, \theta(t)\} = N_c^{-1} \sum_{i=1}^N \xi_i \left[\left\{ \Delta_i - \Lambda_0(X_i) e^{\beta'_w W_i} \right\} \{I(G_i = g) - P(G = g | W_i)\} \right].$$

Note that $M_i \equiv \Delta_i - \Lambda_0(X_i) e^{\beta'_w W_i}$ is the martingale residual under H_0 and thus $U_{\beta_g}\{\alpha, \theta(t)\}$ is essentially the correlation between the martingale residual M_i and $I(G_i = g)$ given W_i within the CCH subcohort. Although it is obvious that $E(M_i | W_i) = 0$ and $\text{Cov}\{M_i, I(G_i = g)\} = 0$ under H_0 , it is interesting that such a covariance is still expected to be 0 within the subcohort which is sampled conditional on Δ_i .

The score statistic involves unknown nuisance parameters $\{\alpha, \theta(t)\}$ which needs to be estimated. We next estimate α as $\tilde{\alpha} = \arg\max_{\alpha} \ell\{\eta = (0, \alpha', \theta^\pi(t))\}$. It is interesting to note that under H_0 , α is the maximizer of the likelihood corresponding to fitting the model (3) to the CCH subset. Thus, under H_0 , the CCH sampling does not introduce bias in estimating α based on the CCH subset only. When W_- is discrete taking unique values $\{w_1, \dots, w_q\}$ and no model assumption is imposed for $P(G = g | W_-)$, the resulting estimator for $\alpha_k = \log\{P(G = g | W_- = w_k)/P(G = 0 | W_- = w_k)\}$ is simply $\alpha_k = \log[\sum_{i \in \Omega_{\text{CCH}}} I(G_i = g, W_{-i} = w_k) / \{\sum_{i \in \Omega_{\text{CCH}}} I(G_i = 0, W_{-i} = w_k)\}]$, for $k = 1, \dots, q$. Let $U_{\beta_G}\{\alpha, \theta^\pi(t)\} = [U_{\beta_1}\{\alpha, \tilde{\theta}^\pi(t)\}, U_{\beta_2}\{\alpha, \tilde{\theta}^\pi(t)\}]'$. We show in Appendix A that

$$\sqrt{N} U_{\beta_G}\{\tilde{\alpha}, \hat{\theta}^\pi(t)\} \sim N(0, \Sigma),$$

where Σ is given in Appendix A. Let $\hat{\Sigma}$ denote a consistent estimator of Σ . We propose to test H_0 with a pseudo score statistic

$$U'_{\beta_G} \{ \tilde{\alpha}, \tilde{\theta}^\pi(t) \} \sum^{-1} U_{\beta_G} \{ \tilde{\alpha}, \hat{\theta}^\pi(t) \},$$

which converges in distribution to a χ^2 distribution with g_0 degrees of freedom under H_0 .

To compute the p-value associated with the pseudo score statistic, we apply a perturbation resampling procedure to obtain $\hat{\Sigma}$. Such a resampling method has been used successfully in the literature to approximating the distribution of regular estimators (e.g. Parzen et al 1994, Jin et al, 2001, Cai et al, 2010). First, let B be a pre-determined number of perturbations and we generate $N_c \times B$ independent and identically distributed non-negative random variables

with mean 1 and variance 1, $\{v_{ib}, i = 1, \dots, N_c, b = 1, \dots, B\}$, where $N_c = \sum_{i=1}^N \xi_i$. For $b = 1, \dots, B$, we can get the perturbed versions of $\hat{\Lambda}_0^\pi(X_i)$ and $\hat{\beta}_w^\pi$, $\hat{\Lambda}_0^\pi(X_i)$ and $\hat{\beta}_w^{(b)}$, by perturbing the influence functions of $\hat{\Lambda}_0^\pi(X_i)$ and $\hat{\beta}_w^\pi$ given in the Appendix A:

$$\hat{\beta}_w^{(b)} = \hat{\beta}_w^\pi + N^{-1} \sum_{i \in \Omega_{CCH}} \hat{\omega}_i \hat{\mathcal{U}}_{\beta_w^\pi}(v_{ib} - 1), \quad \text{and} \quad \hat{\Lambda}_0^{(b)}(t) = \hat{\Lambda}_0^\pi(t) + N^{-1} \sum_{i \in \Omega_{CCH}} \hat{\omega}_i \hat{\mathcal{U}}_{\Lambda_0^\pi}(t)(v_{ib} - 1)$$

where $\hat{\mathcal{U}}_{\beta_w^\pi}$ and $\hat{\mathcal{U}}_{\Lambda_0^\pi}$ are empirical estimates of $\mathcal{U}_{\beta_w^\pi}$ and $\mathcal{U}_{\Lambda_0^\pi}$, respectively. For a given set of $V_b = \{v_{ib}, i = 1, \dots, N_c\}$, we obtain the perturbed counterpart of $\alpha, \tilde{\alpha}^{(b)}$, by maximizing a weighted likelihood corresponding to fitting model (3) to the CCH data with weights V_b . When W_- is discrete taking values $w_1, \dots, w_q$, we may explicitly obtain the perturbed counterpart of α_k as

$\tilde{\alpha}_k^{(b)} = \log \left[\frac{\sum_{i \in \Omega_{CCH}} I(G_i = g, W_{-i} = w_k) v_{ib}}{\sum_{i \in \Omega_{CCH}} I(G_i = 0, W_{-i} = w_k) v_{ib}} \right]$. When parametric models are used to estimate α , $\tilde{\alpha}^{(b)}$ can be obtained as the maximizer of the weighted likelihood function with the contribution from the i subject weighted by V_b . Then the perturbed pseudo score is

$$U_{\beta_g}^{(b)} \{ \tilde{\alpha}^{(b)}, \hat{\theta}^{(b)}(t) \} = \frac{\sum_{i \in \Omega_{CCH}} \left[\{ \Delta_i - \tilde{\Lambda}_0^{(b)}(X_i) e^{W_i' \hat{\beta}_w^{(b)}} \} \{ I(G_i = g) - \mathcal{R}_g(\tilde{\alpha}^{(b)}, W_i) \} \right] v_{ib}}{\sum_{i \in \Omega_{CCH}} v_{ib}}.$$

We take the empirical covariance matrix of the B perturbed pseudo scores, $\hat{\Sigma}$, to approximate the covariance matrix of the observed score statistic.

2.2 A pseudo likelihood method for the estimation of β_G

When interest lies in estimating the model parameters $\eta = (\beta_G', \alpha')'$, we derive a pseudolikelihood estimator, $\hat{\eta}$, as the maximizer of the estimated retrospective likelihood function $\hat{\ell}(\eta) = \ell\{\eta, \hat{\theta}^\pi(t)\}$. Because $\logit\{p(G|X, W)\}$ is not linear in β_G as the ordinary polytomous regression model, we maximize $\hat{\ell}(\eta)$ by using the Newton-Raphson method to solve the following pseudo-likelihood score equation iteratively:

$$U\{\boldsymbol{\eta}; \hat{\boldsymbol{\theta}}^\pi(t)\} = N_c^{-1} \frac{\partial \hat{\ell}(\boldsymbol{\eta})}{\partial \boldsymbol{\eta}} = N_c^{-1} \sum_{i \in \Omega_{\text{CHH}}} U_i\{\boldsymbol{\eta}; \hat{\boldsymbol{\theta}}^\pi(t)\} = 0,$$

$$\text{where } U_i\{\boldsymbol{\eta}; \hat{\boldsymbol{\theta}}^\pi(t)\} = \partial[r_{\mathcal{D}_i} \{G_i | \boldsymbol{\eta}, \hat{\boldsymbol{\theta}}^{(\pi)}(t)\} - \log\{1 + \sum_{g=1}^{g_0} e^{r_{\mathcal{D}_i} \{g | \boldsymbol{\eta}, \hat{\boldsymbol{\theta}}^\pi(t)\}}\}] / \partial \boldsymbol{\eta}.$$

In Appendix B, we show that under suitable regularity conditions, the estimate $\hat{\boldsymbol{\eta}}$ is root-n consistent with

$$N^{\frac{1}{2}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}) = N^{-\frac{1}{2}} \sum_{i \in \Omega_{\text{CCH}}} \mathcal{U}_{\eta i} + o_p(1),$$

where $\mathcal{U}_{\eta i}$ is defined in Appendix B. Therefore, $N^{\frac{1}{2}}(\hat{\boldsymbol{\eta}} - \boldsymbol{\eta})$ converges in distribution to a multivariate normal with mean 0 and variance $\sum_{\boldsymbol{\eta}} = E(\mathcal{U}_{\eta i} \mathcal{U}_{\eta i}' | \xi_i = 1)$. The asymptotic

variance $\Sigma_{\boldsymbol{\eta}}$ can thus be explicitly estimated as $N_c^{-1} \sum_{i \in \Omega_{\text{CCH}}} \hat{\mathcal{U}}_{\eta i} \hat{\mathcal{U}}_{\eta i}'$, where $\hat{\mathcal{U}}_{\eta i}$ is obtained by replacing all theoretical quantities in $\mathcal{U}_{\boldsymbol{\eta}}$ by their empirical counterparts. Alternatively, we may estimate $\Sigma_{\boldsymbol{\eta}}$ via the perturbation similarly as those proposed for Σ .

Specifically, let $\hat{\boldsymbol{\theta}}^{(b)}(t) = \{\hat{\boldsymbol{\beta}}_w^{(b)'}, \hat{\Lambda}_0^{(b)}(t)\}'$ as defined in Section 2.1 and $\boldsymbol{\eta}^{(\hat{b})}$ be the solution to

$$U^{(b)}\{\boldsymbol{\eta}; \hat{\boldsymbol{\theta}}^{(b)}(t)\} = N_c^{-1} \sum_{i \in \Omega_{\text{CCH}}} U_i\{\boldsymbol{\eta}; \hat{\boldsymbol{\theta}}^{(b)}(t)\} = 0.$$

Then $\Sigma_{\boldsymbol{\eta}}$ may be estimated by the empirical variance-covariance matrix of $\{\boldsymbol{\eta}^{(\hat{b})}, b = 1, \dots, B\}$.

3. Data Analysis and Simulation Studies

3.1 Simulation studies

We performed extensive simulation studies to assess the finite sample performance of the proposed methods. Throughout, we let the full cohort size to be $N = 5000$. We generated genotype data G , taking values 0, 1, 2, under Hardy-Weinberg equilibrium. Then we generated a binary W from the logistic regression model $\text{logit } p(W = 1 | G) = \tau_0 + \tau_1 G$. Then T and C were generated from the exponential distributions $\lambda(t | G, W) = \lambda_0 \exp\{\boldsymbol{\beta}_1 I(G = 1) + \boldsymbol{\beta}_2 I(G = 2) + \boldsymbol{\beta}_W W\}$ and $\lambda(c | W) = \lambda_{0c} \exp(\gamma_W W)$, respectively. From the full cohort, we selected a CCH sample consisting of all events and a random sample of the same size. Data on (G, W) is deleted for subjects not selected into the CCH sample. For each configuration, we used 1000 (5000 for calculating empirical type-I error rates) replications to summarize the results. We considered different minor allele frequencies (MAFs) (0.1, 0.2, 0.3), different levels of dependence between G and W ($\tau_1 = 0.2, 0.4, 0.6$), and different effect

sizes ($\beta_1 = \beta_2/2 = 0.2, 0.3, 0.4$). We considered rare, less rare, less common, and common events by varying parameters λ_0 and λ_{0c} . The observed proportion of cases was around 1.0~2.0% for rare events, around 4.0~5.0% for less rare events, around 10.0~12.0% for less common events, and around 20.0~22.0% for common events during the followup period.

We focused on when C depended on W for evaluating the type-I error rate and power of the pseudo score test. We also implemented the IPW method where the Wald statistic was used. All tests were two-sided and at the 0.05 significance level. The results are summarized in Tables 1 and 2. Our pseudo score statistic (“pScore”) maintained the nominal type-I error rate in all scenarios. But the IPW method had seriously inflated type-I error rates when the MAF was small (Table 1) or the event was rare. When MAF = 0.1, the empirical type-I error rate of the IPW method was as high as 0.4, 0.17, 0.1, 0.07 when the event rate was 2%, 5%, 10% and 20%, respectively. When event rate was 2%, the IPW method also has inflated type I error across all scenarios even when MAF = 0.3. As shown in Table 2, our pseudo score test generally has higher empirical power than the IPW for all settings where IPW method is able to maintain the correct size. For instance when the event rate was about 5% and MAF = 0.3, the power for the pseudo score test was around 30% to 40% higher relative to the power for the IPW method. The power improvement appeared to increase with increased association between G and W . It is not appropriate to compare the power at small MAFs or small event rates (2%) because of the inflated type-I error rate of the IPW method as shown in Table 1.

We also examined the finite sample performance of the pseudo likelihood (“pLik”) method in estimating both the genetic association parameters β_G and the parameters in model $P(G = g|W)$, α . We simulated two scenarios with C dependent on W in one and C independent of W in the other. Results are summarized in Tables 3a, 3b and 4. We analyzed each CCH dataset using the pLik method, the IPW method, the augmented inverse probability weighted estimator (AIPW; Kulich and Lin, 2004; Qi et al., 2005), and the maximum likelihood estimator (MLE). For AIPW, the sampling probabilities in the weights were estimated non-parametrically based on $\{\sum_{i=1}^n \xi_i I(W_i=w, \Delta_i=\delta)\} / \{\sum_{i=1}^n I(W_i=w, \Delta_i=\delta)\}$ for $\delta = 0, 1$ and $w = 0, 1$. An EM algorithm was used to obtain the non-parametric MLE. We also analyzed the full cohort data using both standard partial likelihood analysis (“Cox”) and the pLik estimator (pLik_F). The pLik estimator generally had negligible biases across all simulation scenarios, while the IPW, AIPW, and MLE estimators tended to have slightly higher biases when the event was rare. The standard error of the pLik estimator was generally smaller than that of the IPW or AIPW estimators and comparable to that of the MLE. For analyzing the full cohort data, the empirical standard errors for the partial likelihood analysis and the pLik method were nearly identical, further demonstrating the high efficiency of our proposed method. The proposed estimator for the standard errors also performed well, with the average of the estimated standard errors close to the empirical standard errors and empirical coverage probabilities of the confidence intervals close to the nominal level.

3.2 A CCH genetic association study of cardiovascular disease risk in Bangladesh

We applied our method to analyze a case-cohort study nested in the Health Effects of Arsenic Longitudinal Study (Ahsan et al., 2006), which included a random subcohort of 1,375 subjects and 447 incident fatal and nonfatal cases of cardiovascular disease (CVD) from Bangladesh. This study was originally designed to assess interactions between well-water arsenic and a group of selected candidate genes that are related to arsenic metabolism, oxidative stress, inflammation, and endothelial dysfunction on the CVD risk (Wu et al., 2014). We analyzed data from this study to assess the marginal association between each of the 16 SNPs in 10 genes that were previously reported to be associated with the CVD risk, adjusting for baseline variables including sex, age, body mass index (BMI), smoking status, education, systolic blood pressure, and history diabetes. Here, we are interested in assessing whether the presence or absence of the minor allele for each SNP, coding G as 0 or 1, is associated with different CVD risk. We used a logistic regression model for $P(G = 1 | W_-)$, where W_- consists of BMI, smoking status, systolic blood pressure, and history of diabetes. An initial analysis using the IPW method indeed yielded smaller p-values than our pseudo score test. However, it is important to note that the rate of CVD was only about 2% in this study, a setting that our simulation studies suggested drastically inflated type I errors for the IPW method. Therefore, we only report results from our methods as summarized in Table 5. Four SNPs, rs5746088 in gene SOD2, SNP rs2853792 in gene NOS3, rs1789953 in CBS, rs3176871 in VCAM1, were significant using our method. SNP rs5746088 remains significant after Bonferroni correction of multiple comparisons, where the HR comparing presence vs absence of allele A was 0.382 (95% CI: 0.211–0.653). The SOD2 gene was selected in the original study because it was involved in oxidative stress and therefore possibly related to arsenic metabolism. There has been no report yet in the current literature about its relationship with cardiovascular disease risk. Therefore, our result needs to be replicated in independent studies.

4. Discussion

Our methods are designed for assessing the marginal effect of a di-allelic SNP on a time-to-event outcome with CCH data. They are particularly attractive when the genotype and covariate data is available only for subjects in the CCH sample or covariates may be associated with censoring. They require precise followup information only for the CCH sample, which is advantageous when it is not convenient to make such information available for the full cohort. Our methods are naturally extensible to analyze data from stratified CCH designs where cases are supplemented with a random subset selected stratified on factors measured for the full cohort, such as ethnicity and age categories, as long as these sampling variables are fully adjusted for in the Cox regression model. Of course, stratum-specific sampling weights should be used in initial estimates of $\{\hat{\beta}_w^\pi, \hat{\Lambda}_0^\pi(t)\}$.

Results from simulation studies suggested that the statistical efficiency of our proposed method was often comparable to that of NPMLE when the latter is feasible. We note that our methods are not semiparametric efficient, because the distribution of the event time conditional on covariates, $(\cdot, X|W)$, was ignored. One may wonder if the efficiency of our methods can be improved if we use estimators that are more efficient than the IPW method

for obtaining initial parameter estimates. But as evidenced by efficiency comparison between the IPW and AIPW methods (Tables 3 and 4), the low correlation between the genotype variable and covariates implies that efficiency gain is expected to be limited for methods that exploit such correlation to improve efficiency. Therefore, it suffices to use the IPW initial estimator in our methods for such settings.

We note that a similar retrospective model (4) was derived in Prentice and Breslow (Equation 6, 1978), where they discussed the relationship between instantaneous odds ratio for disease development () and the odds of observing exposure (e.g., G in this article). They considered an analysis where cases and controls were matched on follow-up time and therefore maximized a conditional likelihood similar to the partial likelihood. They did not consider analysis of an outcome-dependent sub-sample of the full cohort data. Our work is also related to a pseudo likelihood method in Chen (2001b) for analyzing CCH data with semiparametric transformation models when the follow-up information is available for the full cohort. Adapted to our setting, their approach was based on the joint distribution of genetic and non-genetic variables conditioned on the follow-up time and censoring indicator, in contrast to our approach that is based on the distribution of the genetic variable conditioned on the rest of the variables. Consequently, their pseudo likelihood method involves estimation of the joint distribution of genetic and non-genetic variables, which they proposed to estimate using the random subset of the full cohort that was included as part of the CCH sample. Therefore, our method is attractive when the dimension of non-genetic variables is high. When the sample size is large and the dimension of non-genetic variables is low, the method of Chen (2001b) is feasible. It would be interesting to compare the efficiency of the two methods.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

Shen and Cai were partially supported by grants R01 GM079330, R01 HL089778 and P50 MH106999 from the National Institute of Health. Yang's study was supported by the National Natural Science Foundation of China (No. 11371215). Dr. Chen was supported by National Institute of Environmental Health Sciences (R01-ES016626 and R21-ES020811).

References

- Ahsan H, Chen Y, Parvez F, Argos M, Hussain AI, Momotaj H, Levy D, van Geen A, Howe G, Graziano J. Health Effects of Arsenic Longitudinal Study (HEALS): description of a multidisciplinary epidemiologic investigation. *Journal of Exposure Science and Environmental Epidemiology*. 2006; 16:191–205. [PubMed: 16160703]
- Barlow WE. Robust variance estimation for the CCH design. *Biometrics*. 1994; 50:1064–72. [PubMed: 7786988]
- Bilias Y, Gu M, Ying Z. Towards a general asymptotic theory for Cox model with staggered entry. *Annals of Statistics*. 1997; 25:662–82.
- Borgan O, Langholtz B, Samuelsen SO, Goldstein L, Pogoda J. Exposure stratified CCH designs. *Lifetime Data Analysis*. 2000; 6:39–58. [PubMed: 10763560]
- Breslow NE, Chatterjee N. Design and analysis of two-phase studies with binary outcome applied to Wilms tumour prognosis. *Applied Statistics*. 1999; 48:457–68.

- Breslow NE, Robins JM, Wellner JA. On the semi-parametric efficiency of logistic regression under case-control sampling. *Bernoulli*. 2000; 6:447–55.
- Cai T, Gerds T, Zheng Y, Chen J. Robust Prediction of t-Year Survival with Data from Multiple Studies. *Biometrics*. 2010; 67:436–44. [PubMed: 20670303]
- Christensen LL, Madsen BE, Wikman FP, Wiuf C, Koed K, Tjønneland A, Olsen A, Syvanen AC, Andersen CL, Orntoft TF. The association between genetic variants in hMLH1 and hMSH2 and the development of sporadic colorectal cancer in the Danish population. *BMC Medical Genetics*. 2008; 9:52. [PubMed: 18547406]
- Chen HY, Little RJA. Proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 1999; 94:896–908.
- Chen K, Lo S. Case-cohort and case-control analysis with Cox's model. *Biometrika*. 1999; 86:755–64.
- Chen HY. Analysis of data from the modified case-cohort design by semiparametric transformation regression models. *Biometrika*. 2001a; 88:255–68.
- Chen K. Generalized case-cohort sampling. *Journal of the Royal Statistical Society, series B*. 2001b; 63:791–809.
- Jin Z, Ying Z, Wei LJ. A simple resampling method by perturbing the minimand. *Biometrika*. 2001; 88:381–90.
- Kalbfleisch JD, Lawless JF. Likelihood analysis of multi-state models for disease incidence and mortality. *Statistics in Medicine*. 1988; 1988:146–60.
- Kulich M, Lin DY. Improving the efficiency of relative-risk estimation in CCH studies. *Journal of the American Statistical Association*. 2004; 99:832–44.
- Lin DY, Zeng D. Likelihood-based inference on haplotype effects in genetic association studies. *Journal of the American Statistical Association*. 2006; 101:89–118.
- Nan B, Emond M, Wellner JA. Information bounds for Cox regression models with missing data. *Annals of Statistics*. 2004; 32:723–53.
- Parzen MI, Wei LI, Ying Z. A resampling method based on pivotal estimating functions. *Biometrika*. 1994; 81:341–50.
- Pollard, D. Empirical processes: theory and applications. NSF-CBMS regional conference series in probability and statistics. JSTOR; 1990. p. i-86.
- Prentice RL, Pyke R. Logistic disease incidence models and case-control studies. *Biometrika*. 1979; 66:403–11.
- Prentice R. A CCH design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.
- Prentice RL, Breslow NE. Retrospective studies and failure time model. *Biometrika*. 1978; 65:153–58.
- Qi L, Wang CY, Prentice RL. Weighted estimation for proportional hazards regression with missing covariates. *Journal of the American Statistical Association*. 2005; 100:1250–63.
- Samuelsen SO, Aanestad H, Skrdal A. Stratified CCH analysis of general cohort sampling designs. *Scandinavian journal of statistics*. 2007; 34:103–19.
- Scheike TH, Martinussen T. Maximum likelihood estimation for Cox's regression model under CCH sampling. *Scandinavian Journal of Statistics*. 2004; 31:283–93.
- Self SG, Prentice RL. Asymptotic distribution theory and efficiency results for CCH studies. *The Annals of Statistics*. 1988; 1:64–81.
- Silander K, Alanne M, Kritiansson K, Saarela O, Ripatti S, Auro K, Karvanen J, Kulathinal S, Niemela M, Ellonen P, Vartiainen E, Jousilahti P, Saarela J, Kuulasmaa K, Evans A, Perola M, Salomaa V, Peltonen L. Gender differences in genetic risk profiles for cardiovascular disease. *PLoS One*. 2008; 10:e3615. [PubMed: 18974842]
- Wu F, Jasmine F, Kibriya MG, Liu M, Cheng X, Parvez F, Paul-Brutus R, Paul RR, Sarwar G, Ahmed A, Jiang J, Islam T, Slavkovich V, Rundek T, Demmer RT, Desvarieux M, Ahsan H, Chen Y. Interaction between arsenic exposure from drinking water and genetic susceptibility in carotid intima-media thickness in Bangladesh. *Toxicology and Applied Pharmacology*. 2014; 276:195–203. [PubMed: 24593923]

Table 1

Empirical type-I error rates for the Pseudo score test (pScore) and the Wald test based on the IPW estimator.

MAF	τ_1^*	2% Event Rate		5% Event Rate		10% Event Rate		20% Event Rate	
		pScore	IPW	pScore	IPW	pScore	IPW	pScore	IPW
0.1	0.2	0.039	0.408	0.043	0.176	0.052	0.114	0.053	0.070
	0.4	0.036	0.414	0.039	0.187	0.046	0.103	0.054	0.071
	0.6	0.029	0.424	0.042	0.175	0.046	0.099	0.052	0.080
0.2	0.2	0.056	0.112	0.051	0.067	0.055	0.065	0.047	0.055
	0.4	0.050	0.114	0.051	0.065	0.054	0.065	0.049	0.053
	0.6	0.058	0.107	0.059	0.070	0.052	0.062	0.050	0.055
0.3	0.2	0.059	0.077	0.053	0.068	0.052	0.056	0.052	0.061
	0.4	0.059	0.081	0.053	0.059	0.055	0.056	0.052	0.056
	0.6	0.056	0.074	0.055	0.059	0.051	0.057	0.052	0.054

Power of the Pseudo score test (pScore) and the Wald test based on the IPW estimator when $\beta_2 = 2\beta_1$ for various values of β_1 .

Table 2

(a) Event Rate: 2%						
$\beta_1 = 0.2$			$\beta_1 = 0.3$			$\beta_1 = 0.4$
MAF	τ_1	pScore	IPW	pScore	IPW	pScore
0.1	0.2	0.06	0.22	0.08	0.29	0.17
	0.6	0.06	0.22	0.12	0.28	0.20
0.2	0.2	0.10	0.11	0.19	0.19	0.36
	0.6	0.13	0.12	0.20	0.17	0.33
0.3	0.2	0.12	0.10	0.24	0.18	0.47
	0.6	0.14	0.10	0.28	0.16	0.48

(b) Event Rate: 5%						
$\beta_1 = 0.2$			$\beta_1 = 0.3$			$\beta_1 = 0.4$
MAF	τ_1	pScore	IPW	pScore	IPW	pScore
0.1	0.2	0.12	0.19	0.24	0.29	0.44
	0.6	0.12	0.18	0.25	0.30	0.44
0.2	0.2	0.20	0.16	0.40	0.31	0.72
	0.6	0.21	0.16	0.45	0.31	0.72
0.3	0.2	0.23	0.19	0.55	0.42	0.83
	0.6	0.27	0.17	0.56	0.38	0.86

(c) Event Rate: 10%						
$\beta_1 = 0.2$			$\beta_1 = 0.3$			$\beta_1 = 0.4$
MAF	τ_1	pScore	IPW	pScore	IPW	pScore
0.1	0.2	0.21	0.22	0.45	0.43	0.76
	0.6	0.21	0.21	0.47	0.42	0.76

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

(c) Event Rate: 10%									
MAF	π_1	$\beta_1 = 0.2$		$\beta_1 = 0.3$		$\beta_1 = 0.4$		pScore	IPW
		pScore	IPW	pScore	IPW	pScore	IPW		
0.2	0.2	0.36	0.31	0.74	0.65	0.96	0.96	0.91	
	0.6	0.41	0.29	0.76	0.62	0.95	0.95	0.88	
0.3	0.2	0.49	0.41	0.87	0.79	0.99	0.99	0.97	
	0.6	0.49	0.34	0.88	0.75	0.99	0.99	0.97	
(d) Event Rate: 20%									
MAF	π_1	$\beta_1 = 0.2$		$\beta_1 = 0.3$		$\beta_1 = 0.4$		pScore	IPW
		pScore	IPW	pScore	IPW	pScore	IPW		
0.1	0.2	0.45	0.41	0.81	0.79	0.99	0.99	0.98	
	0.6	0.45	0.38	0.85	0.8	0.98	0.98	0.97	
0.2	0.2	0.74	0.69	0.98	0.96	1	1	1	
	0.6	0.74	0.65	0.98	0.97	1	1	1	
0.3	0.2	0.83	0.77	1	1	1	1	1	
	0.6	0.88	0.79	1	0.99	1	1	1	

Table 3

Bias $\times$ 100, empirical standard error (ESE), average of the estimated standard errors (ASE), and the empirical coverage probabilities of the 95% confidence intervals (CovP) of the proposed pseudo likelihood estimator (pLik), the IPW estimator, the MLE estimator, and the augmented IPW estimator (AIPW) for $\beta_G = (0.2, 0.6)'$. Shown also are results on the bias and ESE based on full cohort comparing the proposed estimator (pLikF) to the maximum partial likelihood estimator (Cox).

(a) Censoring depended on W													
Bias $\times$ 100							ESE _{ASE}						
Event rates	pLik	IPW	MLE	AIPW	pLik _F	Cox	pLik	IPW	MLE	AIPW	pLik _F	Cox	pLik
2%	0.39	1.64	-2.14	-2.95	0.34	0.55	0.31 _{0.30}	0.38 _{0.37}	0.38	0.43	0.23	0.23	94%
	2.35	4.44	-4.47	3.95	-0.71	-1.00	0.49 _{0.50}	0.65 _{0.58}	0.46	0.63	0.31	0.30	96%
5%	-0.39	-0.45	-0.48	-0.64	-0.43	-0.44	0.18 _{0.19}	0.22 _{0.22}	0.13	0.14	0.13	0.13	94%
	0.90	0.71	-2.80	-3.30	-0.65	-0.87	0.29 _{0.30}	0.36 _{0.34}	0.31	0.21	0.20	0.20	94%
10%	0.29	0.27	-0.14	0.33	-0.12	-0.07	0.14 _{0.14}	0.17 _{0.16}	0.13	0.16	0.10	0.10	95%
	-0.53	0.64	0.99	2.59	-0.04	0.54	0.22 _{0.22}	0.25 _{0.25}	0.23	0.25	0.14	0.14	95%
20%	0.19	0.20	0.26	0.61	-0.07	0.06	0.10 _{0.09}	0.11 _{0.10}	0.09	0.09	0.07	0.07	94%
	0.23	-0.39	-0.21	0.82	-0.03	-0.10	0.16 _{0.14}	0.17 _{0.16}	0.14	0.15	0.10	0.10	95%
(b) Censoring was independent of W													
Bias $\times$ 100							ESE _{ASE}						
Event rates	pLik	IPW	MLE	AIPW	pLik _F	Cox	pLik	IPW	MLE	AIPW	pLik _F	Cox	pLik
2%	1.04	2.47	-1.76	2.08	1.14	1.07	0.29 _{0.29}	0.35 _{0.35}	0.29	0.38	0.21	0.21	95%
	1.00	4.41	-3.44	5.62	0.29	-1.81	0.46 _{0.47}	0.58 _{0.55}	0.41	0.57	0.30	0.30	96%
5%	0.21	1.78	-1.52	-0.42	-0.34	-0.39	0.19 _{0.19}	0.24 _{0.22}	0.20	0.23	0.13	0.13	95%
	0.61	0.77	-2.26	1.52	-0.21	-0.37	0.31 _{0.30}	0.32 _{0.35}	0.30	0.36	0.18	0.18	95%
10%	0.05	0.33	-0.29	-0.30	0.29	0.25	0.13 _{0.14}	0.15 _{0.15}	0.13	0.16	0.09	0.10	96%
	-0.31	0.59	0.71	0.87	0.23	0.12	0.22 _{0.21}	0.25 _{0.24}	0.23	0.24	0.14	0.14	96%

(b) Censoring was independent of W

Event rates	Bias×100						ESE _{ASE}						CovP	
	pLik	IPW	MLE	AIPW	pLik _F	Cox	pLik	IPW	MLE	AIPW	pLik _F	Cox	pLik	
20%	-0.02	0.16	0.13	0.27	0.11	0.10	0.10 _{0.09}	0.11 _{0.10}	0.09	0.10	0.06	0.06	95%	
	-0.31	-0.41	-0.10	0.49	0.25	0.18	0.15 _{0.14}	0.17 _{0.16}	0.14	0.16	0.09	0.09	96%	

Table 4

Bias $\times$ 100 and empirical standard error (ESE) for the proposed psuedo likelihood estimator (pLik), the IPW estimator, the MLE estimator, and the augmented IPW estimator (AIPW), for $\alpha = (-0.4122, -0.0122, -2.2449, -1.4449)$.

(a) Censoring depended on W											
Event rates	Bias $\times$ 100				ESE $_{ASE}$				CovP		
	pLik	IPW	MLE	AIPW	pLik	IPW	MLE	AIPW			
2%	0.91	1.59	0.40	1.58	0.28 _{0.29}	0.29	0.26	0.29	95%		
	-0.26	-2.30	1.64	-2.31	0.24 _{0.24}	0.32	0.31	0.32	96%		
	-3.80	4.50	-10.21	5.60	0.54 _{0.59}	0.82	0.61	0.82	97%		
	0.19	-2.32	-4.11	-2.34	0.40 _{0.40}	0.38	0.46	0.38	95%		
5%	0.49	-8.83	-4.61	-6.36	0.17 _{0.18}	0.20	0.16	0.20	96%		
	0.76	4.40	3.17	4.42	0.13 _{0.14}	0.14	0.12	0.14	95%		
	-2.31	6.51	5.65	6.47	0.36 _{0.34}	0.16	0.21	0.16	95%		
	-0.43	0.18	2.62	0.29	0.23 _{0.23}	0.16	0.31	0.16	94%		
10%	-0.81	-0.82	-0.56	-0.71	0.14 _{0.13}	0.16	0.13	0.16	94%		
	-0.24	0.97	1.17	1.01	0.10 _{0.11}	0.11	0.10	0.11	95%		
	-0.04	-1.20	-2.14	-1.42	0.25 _{0.25}	0.30	0.26	0.30	96%		
	-2.21	-0.13	0.31	-0.16	0.17 _{0.17}	0.18	0.17	0.18	94%		
20%	-0.07	-0.62	-0.26	-0.73	0.09 _{0.09}	0.10	0.08	0.10	94%		
	0.03	-0.15	-0.17	-0.23	0.07 _{0.07}	0.07	0.07	0.07	95%		
	-0.53	-0.28	-0.55	-0.06	0.16 _{0.17}	0.19	0.16	0.20	96%		
	-0.87	-0.35	-0.28	-0.49	0.11 _{0.11}	0.11	0.10	0.11	94%		
(b) Censoring was independent of W											
Event rates	Bias $\times$ 100				ESE $_{ASE}$				CovP		
	pLik	IPW	MLE	AIPW	pLik	IPW	MLE	AIPW			
2%	-0.66	-0.42	-2.29	-0.38	0.28 _{0.29}	0.33	0.30	0.33	96%		

(b) Censoring was independent of W

Event rates	Bias×100				ESE _{ASE}				CovP
	pLik	IPW	MLE	AIPW	pLik	IPW	MLE	AIPW	
	-0.60	-0.84	-0.11	-1.00	0.22 _{0.23}	0.25	0.23	0.25	96%
	-4.12	-4.81	-7.89	-4.83	0.55 _{0.59}	0.73	0.62	0.73	96%
	0.60	-2.43	2.00	-2.41	0.34 _{0.37}	0.40	0.38	0.41	95%
5%	-0.16	0.53	-0.63	0.57	0.19 _{0.18}	0.20	0.19	0.20	94%
	-0.35	0.77	1.26	0.86	0.15 _{0.14}	0.16	0.15	0.16	95%
	-1.04	-2.99	-3.04	-2.88	0.35 _{0.35}	0.44	0.36	0.44	94%
10%	0.37	0.14	1.81	0.15	0.22 _{0.23}	0.24	0.24	0.24	95%
	0.50	0.89	0.93	0.91	0.14 _{0.13}	0.15	0.13	0.15	94%
	0.05	0.16	0.14	0.29	0.10 _{0.10}	0.11	0.11	0.11	95%
20%	-0.07	-3.97	-3.71	-4.00	0.25 _{0.25}	0.32	0.25	0.32	95%
	0.26	-1.47	-0.71	-1.42	0.16 _{0.16}	0.17	0.17	0.17	95%
	0.13	0.40	0.47	0.34	0.09 _{0.09}	0.11	0.09	0.11	95%
	0.03	-0.35	-0.29	-0.31	0.06 _{0.07}	0.07	0.07	0.07	95%
	-0.07	-0.34	0.07	-0.39	0.17 _{0.17}	0.21	0.16	0.21	95%
	0.16	-0.99	-0.76	-1.01	0.10 _{0.10}	0.12	0.11	0.12	95%

P-values and hazard ratio (HR) estimates (95% CI) using the proposed pseudo likelihood estimator (pLik) for assessing the association between a set of candidate genes and CVD risk adjusting for clinical variables using data from a CCH study in Bangladesh (Wu et al., 2014).

Table 5

Gene	db SNP ID	Genotype	MAF(%)	P-value	HR(95% CI)
<i>ASSMT</i>	rs10748835	AG+GG vs. AA	A (43.9)	0.081	1.237 (0.951, 1.658)
	rs3740393	CG+CC vs. GG	C (18.9)	0.081	1.233 (0.953, 1.589)
<i>CBS</i>	rs1789953	CT+TT vs. CC	T (23.8)	0.027	0.772 (0.601, 1.005)
	rs234706	GA+AA vs. GG	A (20.9)	0.418	1.112 (0.854, 1.437)
<i>CYBA</i>	rs3794624	GA+AA vs. GG	A (10.7)	0.972	1.004 (0.732, 1.312)
<i>HMOX1</i>	rs2285112	GA+AA vs. GG	A (46.6)	0.443	0.850 (0.664, 1.140)
<i>ICAM1</i>	rs281432	GC+CC vs. GG	C (49.5)	0.286	0.846 (0.637, 1.117)
<i>IL6</i>	rs2069832	GA+AA vs. GG	A (11.9)	0.249	0.833 (0.610, 1.158)
<i>MTHFR</i>	rs1801133	CT+TT vs. CC	T (10.9)	0.088	1.267 (0.945, 1.662)
<i>NOS3</i>	rs2853792	AG+GG vs. AA	G (31.0)	0.007	0.702 (0.547, 0.901)
<i>SOD2</i>	rs5746088	GA+AA vs. GG	A (5.2)	< 10 ⁻⁴	0.382 (0.211, 0.653)
<i>VCAM1</i>	rs5746136	GA+AA vs. GG	A (30.6)	0.182	1.174 (0.927, 1.480)
	rs2392221	CT+TT vs. CC	T (17.3)	0.405	0.866 (0.670, 1.138)
	rs3176860	GA+AA vs. GG	A (43.3)	0.355	1.157 (0.887, 1.522)
	rs3176870	GA+AA vs. GG	A (8.7)	0.054	1.333 (0.967, 1.794)
	rs3176871	GA+AA vs. GG	A (5.1)	0.031	1.532 (1.021, 2.284)