HHS Public Access

Author manuscript

Lifetime Data Anal. Author manuscript; available in PMC 2019 July 01.

Published in final edited form as:

Lifetime Data Anal. 2018 July ; 24(3): 443–463. doi:10.1007/s10985-017-9402-7.

A Regularized Variable Selection Procedure in Additive Hazards Model with Stratified Case-Cohort Design

Ai Ni and

Department of Epidemiology and Biostatistics, Memorial Sloan Kettering Cancer Center, 485 Lexington Ave., New York, NY, USA, 10017

Jianwen Cai

Department of Biostatistics, University of North Carolina at Chapel Hill, 135 Dauer Drive, Chapel Hill, NC, USA, 27599

Abstract

Case-cohort designs are commonly used in large epidemiological studies to reduce the cost associated with covariate measurement. In many such studies the number of covariates is very large. An efficient variable selection method is needed for case-cohort studies where the covariates are only observed in a subset of the sample. Current literature on this topic has been focused on the proportional hazards model. However, in many studies the additive hazards model is preferred over the proportional hazards model either because the proportional hazards assumption is violated or the additive hazards model provides more relevent information to the research question. Motivated by one such study, the Atherosclerosis Risk in Communities (ARIC) study, we investigate the properties of a regularized variable selection procedure in stratified case-cohort design under an additive hazards model with a diverging number of parameters. We establish the consistency and asymptotic normality of the penalized estimator and prove its oracle property. Simulation studies are conducted to assess the finite sample performance of the proposed method with a modified cross-validation tuning parameter selection methods. We apply the variable selection procedure to the ARIC study to demonstrate its practical use.

Keywords

Additive Hazards Model; Diverging Number of Parameters; SCAD; Stratified Case-Cohort Design; Survival Analysis; Variable Selection

1 Introduction

In large-scale epidemiological cohort studies, investigators are usually interested in assessing the association between a time-to-event outcome and a large number of risk factors. Collecting information on risk factors often requires expensive bioassays and precious biological specimens such as serum and genetic material. Prentice (1986) proposed a case-cohort design to reduce the cost and effort in measuring expensive covariates without

Web appendix

Additional supporting information may be found in the Web appendix available at the journal's website.

decreasing much efficiency in the estimation. In a case-cohort design, the complete covariate information is only obtained from a randomly sampled subset of the full cohort plus all subjects who developed the outcome event. In practice, some covariates that are correlated with the more expensive exposure variables may be readily available for the entire cohort. One example is the Atherosclerosis Risk in Communities (ARIC) study (Ballantyne et al., 2004), where a cohort of 15,792 participants 45 to 64 years old were sampled from four U.S. communities and were followed for ten years for the development of Coronary Heart Disease (CHD). The primary interest was to assess the association between the protein hs-CRP level and the risk of CHD. To preserve stored plasma and reduce costs, it is desirable to only measure the hs-CRP on a subset of the entire cohort. On the other hand, sex, race, and baseline age were available for all participants. To utilize the fully observed covariates to gain estimation efficiency, Borgan et al. (2000) proposed a stratified case-cohort design where the strata are defined by these covariates. The ARIC study implemented this stratified case-cohort design with stratification on sex, race, and baseline age. The hs-CRP level was measured only on the stratum-specific random subsets plus all incident CHD cases.

Cox proportional hazards model (Cox, 1972) is commonly used for the analysis of time-to-event data. However, the critical assumption of proportional hazards may fail to hold in many situations, making the Cox model invalid. For example, in the ARIC study there is evidence that the hazard of CHD does not satisfy the proportionality assumption (Kang et al., 2013). Even if the proportional hazards assumption is satisfied, investigators are sometimes more interested in the absolute hazard difference as a measure of covariate effect because it is more relevant to public health. Under rare event assumption, which is true for many case-cohort studies, the cumulative hazard difference approximates the attributable risk (the difference in the event rate per unit change in the exposure variable), which translates directly into the number of events that would be avoided by eliminating a particular exposure. Moreover, the risk difference is easier to interpret and communicate to medical practitioners. Therefore, the additive hazards model is often used as an important alternative to the Cox proportional hazards model. As its name suggests, the additive hazards model assumes that the effect of covariates on the risk of event is additive. Since Aalen (1980) first introduced the additive hazards model, many authors have investigated its estimation procedure and the properties of the estimator. Lin & Ying (1994) proposed a semiparametric estimating equation for a special case of additive hazards model where the regression coefficients are time-independent. The authors derived the limiting distribution of the estimator and studied its semiparametric efficiency. Kulich & Lin (2000) extended this estimation method to case-cohort design and assessed its asymptotic relative efficiency with respect to the full cohort analysis.

In case-cohort studies where a large number of covariates are collected, researchers are often interested in selecting a subset of the covariates that are related to the event of interest. With the inclusion of interaction and polynomial terms, the number of candidate covariates can be very large. In the ARIC study, there are a number of potential confounders or effect modifiers that need to be considered in the modeling process. With the pairwise interactions between hs-CRP level and all the other covariates as well as the squared continuous covariates, the total number of candidate covariates becomes quite large in comparison to the number of events. As Huber (1973) argued, in the context of variable selection the number

of parameters should be considered as increasing with sample size. Ni et al. (2016) developed a regularized variable selection method for a case-cohort design under Cox's proportional hazards model with a diverging number of parameters. Such a method needs to be developed for studies such as the ARIC study where the additive hazards model is used under a case-cohort design.

Regularized variable selection procedures have gained much success over the last few decades. Under certain regularity conditions, these procedures can simultaneously select variables and estimate their coefficients. Among various penalty functions used in these procedures, the smoothly clipped absolute deviation (SCAD) penalty (Fan & Li, 2001) and a few others have been shown to possess the oracle property (Fan & Li, 2001). The SCAD variable selection procedure has been applied to the additive hazards model (Lin & Lv, 2013). However, it has not been investigated under a stratified case-cohort design, which is the objective of this paper. Tuning parameter selection is critical for the success of regularized variable selection. We also propose a modified cross-validation based tuning parameter selection strategy to overcome the issue of overfitting with the conventional cross-validation method, and empirically evaluate its performance under large cohort sizes and high censoring rates, which are two typical features of case-cohort studies.

2 Additive Hazards Model with A Stratified Case-Cohort Design

Suppose the full cohort of size n is divided into S mutually exclusive strata based on some categorical variables that are available for all subjects. In this paper we assume S is finite. For subject i in stratum s, let $Z_{si}(t)$ be the $d_n \times 1$ possibly time-dependent covariate vector. We allow d_n to increase with n but at a slower rate that will be determined later. Without loss of generality, we partition the real-valued true parameter vector β_0 as $(\beta_{0,I}^T, \beta_{0,II}^T)^T$, where $\beta_{0,I}$ and $\beta_{0,II}$ are the nonzero and zero components of β_0, respectively. Denote by k_n the dimension of $\beta_{0,I}$, which is also allowed to diverge with n and k_n/d_n converges to a constant $c \in [0, 1]$. Although the dimensions of the true parameter and covariates all depend on n, we suppress the subscript n for notational simplicity.

Let T_{si} and C_{si} be respectively the time to the outcome event and the censoring time for subject i in stratum s, which are independent conditional on Z_{si}. Let $X_{si} = \min(T_{si}, C_{si})$ be the observed time and $\Delta_{si} = I(T_{si} \le C_{si})$ be the censoring indicator, where $I(\cdot)$ is an indicator function. Let τ be the time at the end of study. Define the counting process $N_{si}(t) = I(X_{si} \le t, \Delta_{si} = 1)$, and the at risk process $Y_{si}(t) = I(X_{si} \ge t)$. Let $\lambda_{si}(t)$ denote the hazard function for subject i in stratum s. The additive hazards model assumes $h_{si}\{t|Z_{si}(t)\} = h_0(t) + \beta_0^T Z_{si}(t)$, where $h_0(t)$ is a common baseline hazard function for all strata, and β_0 is constant over time. Under the stratified case-cohort design proposed in Borgan et al. (2000), we randomly select a subcohort of fixed size from each stratum. We assume that the selection of subcohort is independent across the strata. Let $\tilde{n}_s$ denote the subcohort size in stratum s with size n_s, and ξ_{si} be the indicator of subject i being selected into the subcohort in stratum s. Then for subject in stratum $s = 1, \ldots, S$, the selection probability $\text{pr}(\xi_{si} = 1) = \tilde{n}_s/n_s = a_s$. Under simple random sampling $(\xi_{s1}, \ldots, \xi_{sn_s})$ are correlated. The cases (i.e. individuals who developed the event) in each stratum that are not selected into the corresponding subcohort

are added to it to form the stratum-specific case-cohort samples. Assuming the censoring time is available for the noncases outside the subcohorts and complete covariate history is available for cases outside the subcohorts, we consider the following estimating function for β_0,

$$U_n(\beta_n) = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \int_0^{\tau} \rho_{si}(t) \left\{ Z_{si}(t) - \widetilde{Z}(t) \right\} \{ dN_{si}(t) - Y_{si}(t) \beta_n^T Z_{si}(t) dt \},$$

where

$\widetilde{Z}(t) = \sum_{s=1}^{S} \sum_{j=1}^{n_s} \rho_{sj}(t) Y_{sj}(t) Z_{sj}(t) / \sum_{s=1}^{S} \sum_{j=1}^{n_s} \rho_{sj}(t) Y_{sj}(t)$, $\rho_{si}(t) = \Delta_{si} + (1 - \Delta_{si}) \xi_{si} \hat{\alpha}_s^{-1}(t)$, and $\hat{\alpha}_s(t) = \sum_{i=1}^{n_s} \xi_{si} (1 - \Delta_{si}) Y_{si}(t) / \sum_{i=1}^{n_s} (1 - \Delta_{si}) Y_{si}(t)$. This estimating equation is based on Kulich & Lin (2000) with the selection probability α_s replaced by its time-dependent sample estimate $\hat{\alpha}_s(t)$. The estimator $\tilde{\beta}_n$ solves $U_n(\beta_n)$ and takes on a closed form

$$\widetilde{\beta}_n = \left[\sum_{s=1}^{S} \sum_{i=1}^{n_s} \int_0^{\tau} \rho_{si}(t) \left\{ Z_{si}(t) - \widetilde{Z}(t) \right\}^{\otimes 2} Y_{si}(t) dt \right]^{-1} \times \quad (1)$$

$$\left[\sum_{s=1}^{S} \sum_{i=1}^{n_s} \int_0^{\tau} \left\{ Z_{si}(t) - \widetilde{Z}(t) \right\} dN_{si}(t) \right],$$

where $a^{\otimes 2} = aa^T$ for a vector a.

3 Variable Selection in Additive Hazards Model with A Case-Cohort Design

3.1 Penalized loss function

Unlike the Cox proportional hazards model where the log-partial likelihood function is a natural choice of loss function for variable selection, under additive hazards model the likelihood function is difficult to work with due to the nonparametric estimate of the baseline hazard function and the additive structure. Motivated by the similarity between the Lin-Ying estimator for additive hazards model (Lin & Ying, 1994) and the least square estimator, Martinussen & Scheike (2009) proposed a loss function that is the integral of the Lin-Ying estimating equation with respect to β_n. Similarly, we propose a loss function under stratified case-cohort design

$$\widetilde{L}_n(\beta_n) = \frac{1}{2} (\beta_n^T \widetilde{A}_n \beta_n - 2 \beta_n^T \widetilde{b}_n),$$

where

$$\widetilde{A}_n = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \int_0^{\tau} \rho_{si}(t)\{Z_{si} - \widetilde{Z}(t)\}^{\otimes 2} Y_{si}(t)dt,$$

$$\widetilde{b}_n = \sum_{s=1}^{S} \sum_{i=1}^{n_s} \int_0^{\tau} \{Z_{si} - \widetilde{Z}(t)\} dN_{si}(t).$$

We then propose the following objective function for variable selection,

$$\widetilde{Q}_n(\beta_n) = \widetilde{L}_n(\beta_n) + \sum_{j=1}^{d_n} P_{\lambda_{nj}}(|\beta_{nj}|), \quad (2)$$

where $P_{\lambda_{nj}}(\cdot)$ is a nonnegative penalty function with $P_{\lambda_{nj}}(0) = 0$. The tuning parameter λ_{nj} controlling the magnitude of the penalty. We use SCAD penalty with covariate-specific λ_{nj}. When $\lambda_{nj} = 0$, $P_{\lambda_{nj}}(|\beta_{nj}|) = 0$. The first derivative of the SCAD penalty is $P'_{\lambda_{nj}}(\theta) = \lambda_{nj} I(\theta \le \lambda_{nj}) + (a\lambda_{nj} - \theta)_+ (a-1)^{-1} I(\theta > \lambda_{nj})$ for some $a > 2$ and $\theta > 0$.

3.2 Asymptotic Properties of the Penalized Estimator

Denote the penalized estimator that minimizes (2) as $\hat{\beta}_n = (\hat{\beta}_{n,I}^T, \hat{\beta}_{n,II}^T)^T$, where $\hat{\beta}_{n,I}$ and $\hat{\beta}_{n,II}$ are the penalized estimators of $\beta_{0,I}$ and $\beta_{0,II}$, respectively. Let $a^{\otimes 0} = 1$, $a^{\otimes 1} = a$, and $a^{\otimes 2} = aa^T$ for a vector a. Let $P'_{\lambda}(|\beta|) = \partial P_{\lambda}(|\beta|)/\partial\beta$ and $P''_{\lambda}(|\beta|) = \partial^2 P_{\lambda}(|\beta|)/\partial\beta^2$. We first define the following notations for each n.

$$S_n^{(k)}(t) = n^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n_s} Y_{si}(t) Z_{si}(t)^{\otimes k}, \quad k = 0, 1, 2,$$

$$\widetilde{S}_n^{(k)}(t) = n^{-1} \sum_{s=1}^{S} \sum_{i=1}^{n_s} \rho_{si}(t) Y_{si}(t) Z_{si}(t)^{\otimes k}, \quad k = 0, 1, 2,$$

$$s_n^{(k)}(t) = E\{S_n^{(k)}(t)\}, k = 0, 1, 2, \qquad e_n(t) = \frac{s_n^{(1)}(t)}{s_n^{(0)}(t)},$$

$$\mathscr{A}_n = E\left[\int_0^\tau \{Z(t) - e_n(t)\}^{\otimes 2} Y(t)dt\right], \qquad \Gamma_n(\beta_n) = \mathrm{Var}\left\{n^{-1/2}\frac{\partial \tilde{L}_n(\beta_n)}{\partial \beta_n}\right\},$$

$$\phi_n = \max_{1 \le j \le k_n} \{|P'_{\lambda_{nj}}(|\beta_{0j}|)|\}, \qquad \psi_n = \max_{1 \le j \le k_n} \{|P''_{\lambda_{nj}}(|\beta_{0j}|)|\},$$

$$\Psi_n = \mathrm{diag}\{P''_{\lambda_{n1}}(|\beta_{01}|), \ldots, P''_{\lambda_{nk_n}}(|\beta_{0k_n}|)\},$$

$$\Phi_n = \{P'_{\lambda_{n1}}(|\beta_{01}|)\mathrm{sgn}(\beta_{01}), \ldots, P'_{\lambda_{nk_n}}(|\beta_{0k_n}|)\mathrm{sgn}(\beta_{0k_n})\}^T.$$

Only main theorems are presented in this section. Since the integrands of $\tilde{A}_n$ and $\tilde{b}_n$ involves Δ_{si} which is not predictable with respect to the filtration generated by $Y_{si}(t)$, $N_{si}(t)$, and $Z_{si}(t)$, the standard martingale convergence theorem cannot be used to establish the asymptotic results. We instead use empirical process techniques in the proof. The regularity conditions and the outline of the proofs are provided in Web appendix. We first establish the consistency of the penalized estimator and establish its convergence rate.

Theorem 1—Under Conditions (A) to (C) in Web appendix, if $\psi_n \to 0$ and $d_n^2/n \to 0$ as $n \to \infty$, then with probability tending to one there exists a local minimizer $\hat{\beta}_n$ of $\tilde{Q}_n(\beta_n)$, as defined in (2), such that $\|\hat{\beta}_n - \beta_0\| = O_p\{d_n^{1/2}(n^{-1/2} + \phi_n)\}$.

From Theorem 1 one can obtain a $n^{1/2}d_n^{-1/2}$-consistent penalized estimator provided that $\phi_n = O(n^{-1/2})$, which is the case for SCAD penalty. The following theorem establishes the oracle property of the consistent penalized estimator.

Theorem 2—Under Conditions (A) to (E) in Web appendix, if $\psi_n \to 0$, $d_n^2/n \to 0$, $\lambda_{nj} \to 0$, $\lambda_{nj}n^{1/2}d_n^{-1/2} \to \infty$ for $j = 1, \ldots, d_n$, and $\phi_n = O(n^{-1/2})$ as $n \to \infty$, then the $n^{1/2}d_n^{-1/2}$-consistent local minimizer $\hat{\beta}_n = (\hat{\beta}_{n,I}^T, \hat{\beta}_{n,II}^T)^T$ must satisfy (i) $\hat{\beta}_{n,II} = 0$ with probability tending to one, and (ii) for any nonzero $k_n \times 1$ constant vector u_n with $\|u_n\| = 1$,

$$n^{1/2}u_n^T\Gamma_{n11}^{-1/2}(\beta_0)(\mathscr{A}_{n11} + \Psi_n)\{\hat{\beta}_{n,I} - \beta_{0,I} + (\mathscr{A}_{n11} + \Psi_n)^{-1}\Phi_n\} \to N(0,1)$$

in distribution, where $\mathscr{A}_{n11}$ consists of the first $k_n \times k_n$ components of $\mathscr{A}_n$, and $\Gamma_{n11}(\beta_0)$ consists of the first $k_n \times k_n$ components of $\Gamma_n(\beta_0)$.

For the SCAD penalty, $\phi_n = 0$, $\Psi_n = 0$, and $\Phi_n = 0$ for large n under Condition (E) in in Web appendix, and the result of Theorem 2 reduces to

$$n^{1/2} u_n^T \Gamma_{n11}^{-1/2}(\beta_0) \mathcal{A}_{n11} (\hat{\beta}_{n,I} - \beta_{0,I}) \to N(0,1)$$

in distribution.

4 Considerations in Practical Implementation

4.1 Selection of Tuning Parameters

The tuning parameter λ_n involved in the SCAD penalty function controls the complexity of the selected model. A sequence of increasing λ_n values gives rise to a solution path containing models with decreasing dimension. The oracle property of the variable selection procedure guarantees that the solution path contains the true model. In practice, however, one has to choose λ_n to identify the true model. The typical methods of tuning parameter selection are data-driven procedures such as K-fold cross-validation and generalized cross-validation (GCV) (Craven & Wahba, 1979). The d_n-dimensional optimization problem is difficult to solve in practice. We follow Ni et al. (2016) to take $\lambda_{nj} = \lambda_n \widehat{\mathrm{SE}}(\beta_{nj}^{(0)})$, where $\widehat{\mathrm{SE}}(\beta_{nj}^{(0)})$ is the estimated standard error of the unpenalized estimator. Then the optimization problem reduces to one-dimensional and a grid-search can be performed. In the literature of variable selection in Cox's proportional hazards model the GCV is predominantly used due to the availability of the partial likelihood function. Under additive hazards model, however, no such likelihood function is available. Therefore, we use five-fold cross-validation with $\tilde{L}_n(\beta_n)$ as a natural choice of loss function. Denote the full dataset by D and the training and validation dataset by $D - D^\nu$ and D^ν, respectively, for $\nu = 1, \ldots, 5$. For each λ_n, compute $\tilde{L}_{n^\nu}\{\hat{\beta}_n^{-\nu}(\lambda_n)\}$ based on the validation dataset, where n^ν is the sample size of dataset D^ν and $\hat{\beta}_n^{-\nu}(\lambda_n)$ is the penalized estimate based on the training dataset and λ_n. The conventional cross-validation statistics is defined as

$$\mathrm{CV}(\lambda_n) = \sum_{\nu=1}^{5} \tilde{L}_{n^\nu}\{\hat{\beta}_n^{-\nu}(\lambda_n)\}, \quad (3)$$

and λ_n is chosen by minimizing (3). Since the cross-validation method aims at minimizing the prediction error rather than model selection consistency, it tends to overfit the model (Hastie et al., 2009). We propose a modified cross-validation method that incorporates an additional penalty on the model size in the cross-validation statistics. The modified statistic is defined as

$$\mathrm{CV}^{\mathrm{P}}(\lambda_n) = \sum_{\nu=1}^{5} [\tilde{L}_{n^\nu}\{\hat{\beta}_n^{-\nu}(\lambda_n)\} + k^{-\nu}], \quad (4)$$

where $k^{-\nu}$ is the number of nonzero components of $\hat{\beta}_n^{-\nu}$. The penalized loss function in (4) is analogous to the Akaike information criterion (AIC) (Akaike, 1973). Thus, the proposed statistic can be seen as a combination of cross-validation and AIC. We denote the minimizer of (3) and (4) as λ_n^{CV} and λ_n^{CVP}, respectively. In the simulation section, we empirically investigate the model selection performance of these two tuning parameter selection criteria. According to Fan & Li (2001), the second tuning parameter a in the SCAD penalty is set to 3.7.

4.2 Estimation Procedure

Since the SCAD penalty function is singular at the origin, in practical implementation the penalized estimator cannot be directly obtained by solving the first derivative of (2). Instead, we follow Fan & Li (2001) to use a local quadratic approximation (LQA) to the penalty function. The unpenalized loss function $\tilde{L}_n(\beta_n)$ is a special case of (2) with $P_{\lambda_{nj}}(|\beta_{nj}|) = 0$ for all $j = 1, \ldots, d_n$. Applying Theorem 1 with $\phi_n = 0$, we know there exists a $n^{1/2}d_n^{-1/2}$-consistent minimizer of (2). The concavity of (2) ensures that the minimizer is unique. This minimizer $\hat{\beta}_n^{(0)}$ is used as the initial value for the LQA algorithm. A sequence of about 40 increasing values of the tuning parameter denoted as $\Lambda = \{\lambda_1, \ldots, \lambda_{40}\}$ needs to be specified before the procedure. Based on our simulation experience, the range rarely exceeds [0, 3] and the estimation result is insensitive to the fineness of the grid. The iterative estimation procedure is summarized as follows.

Step 1 Randomly split data into five equal-size partitions;

Step 2 Use four fifth of the data to fit an unpenalized stratified additive hazards regression with all candidate covariates to obtain the initial value $\hat{\beta}_n^{(0)}$ and $\widehat{\mathrm{SE}}(\hat{\beta}_n^{(0)})$ using (1);

Step 3 For a chosen tuning parameter λ_i, set $\lambda_{nj} = \widehat{\mathrm{SE}}(\hat{\beta}_{nj}^{(0)})\lambda_i$ and constant $c_j = \lambda_{nj}$ for $j = 1, \ldots, d_n$;

Step 4 At any iteration k, for $j = 1, \ldots, d_n$, if $|\hat{\beta}_{nj}^{(k)}| < c_j$ then set $\hat{\beta}_{nj}^{(k+1)} = 0$. Otherwise, the SCAD penalty is approximated by a quadratic function as

$$P_{\lambda_{nj}}(|\beta_{nj}|) \approx P_{\lambda_{nj}}(|\hat{\beta}_{nj}^{(k)}|) + \frac{P'_{\lambda_{nj}}(|\hat{\beta}_{nj}^{(k)}|)}{2|\hat{\beta}_{nj}^{(k)}|}\left\{\beta_{nj}^2 - (\hat{\beta}_{nj}^{(k)})^2\right\},$$

and therefore $P'_{\lambda_{nj}}(|\beta_{nj}|) \approx \{P'_{\lambda_{nj}}(|\hat{\beta}_{nj}^{(k)}|)/|\hat{\beta}_{nj}^{(k)}|\}\beta_{nj}$. Then fit a penalized stratified additive hazards regression with covariates whose $|\hat{\beta}_{nj}^{(k)}| \geq c_j$ and the objective function (2) with the above quadratic penalty to obtain the nonzero components of $\hat{\beta}_n^{(k+1)}$. This is essentially a ridge regression and a closed form

is available for $\hat{\beta}_n^{(k+1)}$. Denote the nonzero component of $\hat{\beta}_n^{(k+1)}$ as $\hat{\beta}_n^*$. The sandwich estimate of the covariance matrix of $\hat{\beta}_n^*$ can be obtained as

$$\widehat{\mathrm{Cov}}(\hat{\beta}_n^*) = \left\{ \frac{\partial^2 \widetilde{L}_n(\hat{\beta}_n^*)}{\partial (\hat{\beta}_n^*)^2} + n\Phi_n(\hat{\beta}_n^*) \right\}^{-1} \widehat{\mathrm{Var}} \left\{ \frac{\partial \widetilde{L}_n(\hat{\beta}_n^*)}{\partial \hat{\beta}_n^*} \right\} \times \left\{ \frac{\partial^2 \widetilde{L}_n(\hat{\beta}_n^*)}{\partial (\hat{\beta}_n^*)^2} + n\Phi_n(\hat{\beta}_n^*) \right\}^{-1}$$

$$= \{\widetilde{A}_n^* + n\Phi_n(\hat{\beta}_n^*)\}^{-1} n\hat{\Gamma}_n(\hat{\beta}_n^*)\}\{\{\widetilde{A}_n^* + n\Phi_n(\hat{\beta}_n^*)\}^{-1},$$

where $\widetilde{A}_n^*$ is the sub-matrix of $\tilde{A}_n$ corresponding to $\hat{\beta}_n^*$, $\hat{\Gamma}_n(\hat{\beta}_n^*)$ is the estimate of $\Gamma_n(\hat{\beta}_n^*)$, $\Phi_n(\hat{\beta}_n^*) = \mathrm{diag}\{P'_{\lambda_{n1}}(|\hat{\beta}_{n1}^*|)/|\hat{\beta}_{n1}^*|, \ldots, P'_{\lambda_{nk_n^*}}(|\hat{\beta}_{nk_n^*}^*|)/|\hat{\beta}_{nk_n^*}^*|\}$, and k_n^* is the dimension of $\hat{\beta}_n^*$. The sandwich estimate of the covariance matrix does not apply to the zero estimate of the parameters;

Step 5 Iterate the above step until convergence or no nonzero parameter estimate is left. Then apply the final model to the remaining one fifth of data to compute one summand in (4);

Step 6 Repeat Step 2 to 5 rotating over the five data partitions to complete the five-fold cross-validation, and compute $\mathrm{CV}^{\mathrm{P}}(\lambda_i)$ using (4);

Step 7 Repeat Step 1 to 6 over all values in Λ and choose the tuning parameter λ_0 with the smallest CV^{P}. Then use λ_0 and the full dataset to fit a penalized stratified additive hazards model to obtain the final estimate $\hat{\beta}_n$ and $\widehat{\mathrm{SE}}(\hat{\beta}_n)$.

5 Simulation Studies

5.1 Simulation Setup

Independent failure times are generated by the additive hazards model $h_{si}(t) = h_0(t) + \beta_0^T Z_{si}(t)$ under the constraint $h_{si}(t) \geq 0$ ($s = 1, 2$; $i = 1, \ldots, n_s$). We set $h_0(t) = 2$ and the dimension of β_0 to be $d_n = [0.5 * n_c^{1/2 - 1/500}]$ to reflect its dependence on sample size, where n_c is the number of cases and $[x]$ rounds x to the nearest integer. We use n_c instead of n to determine the model size because the former better represents the amount of information in the dataset. The smallest nonzero parameter in terms of the absolute value is set to 0.70 or 0.43, which represents 35% and 22% increase from the baseline hazard for one standard deviation increase in the covariate. The remaining nonzero parameters recycling from values –0.8 and 1. There is one nonzero parameter for every two zero parameters. To generate the design matrix and strata, we first generate a $(d_n + 1)$-dimensional multivariate standard normal variable Z^* with $\mathrm{corr}(Z_i^*, Z_j^*) = 0.5^{|i-j|}$. The first component is then dichotomized with a cutoff value of zero and used to define two strata. For the remaining d_n components, we dichotomize half of them with a cutoff value of zero. As a result, the design matrix consists

of a mixture of correlated binary and continuous covariates that are correlated with the stratification variable. A simple random sample is selected independently from each stratum with the same sampling probability. Censoring times C_i are generated from a uniform distribution U(0, c) where c is adjusted to achieve desired censoring percentage.

Two sample sizes, two censoring rates, and two sampling probabilities of the random subcohort are considered for each minimum effect size ($\beta_{\min}$=0.70 and 0.43). The sampling probabilities are chosen so that the case to noncase ratio in the case-cohort sample is 1:1 or 1:2. Comparisons are made on the performance of penalized variable selection procedures with tuning parameter λ_n^{CV} and λ_n^{CVP}. We include the backwards elimination as a competing variable selection method. We also include as a benchmark the oracle procedure where the correct subset of covariates is used to fit the model. As the censoring rate in case-cohort studies is typically high, we set it to 85% and 90% in our simulation to better mimic real-world studies. For each setting 500 replications are conducted.

The performance of the model selection procedure is evaluated by model error defined as $\mathrm{ME}(\hat{\mu}) = E\{E(Y|Z) - \hat{\mu}(Z)\}^2$. Under the additive hazards model with constant baseline hazard h_0, it can be shown that $E(Y|Z) = (h_0 + \beta_0^T Z)^{-1}$ and $\hat{\mu}(Z) = (h_0 + \hat{\beta}_n^T Z)^{-1}$. Therefore, $\mathrm{ME}(\hat{\mu}) = E\{(h_0 + \hat{\beta}_n^T Z)^{-1} - (h_0 + \beta_0^T Z)^{-1}\}^2$. We further define the relative model error (RME) of a model selection procedure as the ratio of its model error to that of the unpenalized estimates from the full model. Following Tibshirani (1996), we use the median and the median absolute deviation (MAD) of the relative model error to compare the performance of different model selection procedures. We also calculate the average number of zero parameters incorrectly estimated as nonzero (FP), the average number of nonzero parameters incorrectly estimated as zero (FN), the median model size (MS), and the overall rate of identifying the true model (RITM). In addition, point estimates, empirical and model-based standard errors, and the 95% coverage are calculated for $\hat{\beta}_{\min}$ using replications with nonzero $\hat{\beta}_{\min}$.

5.2 Simulation Results

Table 1 summarizes the model selection performance when $\beta_{\min}$ = 0.70. The CVP tuning parameter selection method outperforms the CV tuning parameter selection method in all settings in terms of relative model error (RME) and the rate of identifying the true model (RITM). It also outperforms the backwards elimination method except for the scenarios with n = 10000 and R = 1:1. Higher sampling proportion and lower censoring rate are associated with better model selection performance of the CVP and CV methods despite the associated larger number of parameters but seem to adversely affect the performance of the backwards elimination method. This suggests that the performance of the latter method is more sensitive to the number of parameters. Compared to the CVP method, the CV and backwards elimination methods tend to overfit the model as shown by the FP columns. Table 2 summarizes the estimation result of $\beta_{\min}$ under settings in Table 1. Given that $\beta_{\min}$ is correctly identified as nonzero, all procedures produce reasonably unbiased point estimates. The model-based standard error estimates are very close to the empirical standard errors and the 95% coverage is close to the nominal level.

Table 3 summarizes the model selection performance when $\beta_{\min} = 0.43$. Under the same setting, there is a decrease in the model selection performance for all three procedures in comparison to that with larger $\beta_{\min}$. This is expected as smaller effect is more difficult to detect. Nevertheless, similar to Table 1, the CVP method outperforms the CV method in all settings. It also outperforms the backwards elimination method under 85% censoring rate but performs slightly worse than it under 90% censoring rate. However, since the performance of the backwards elimination method is more strongly affected by the number of parameters than is the CVP method as shown in Table 1 and 3, overall it seems that the regularized variable selection with CVP tuning parameter selection method is a better strategy than the backwards elimination. Table 4 shows the estimation result of $\beta_{\min}$ under settings in Table 3. Conditional on correctly identifying $\beta_{\min}$ all three procedures produce noticeable overestimation in the parameter and its standard error. The bias diminishes as the rate of identifying true model increases, suggesting that the bias is likely due to the fact that zero estimates are excluded from the calculation. We also performed inference using all simulation replications and the point estimates are substantially smaller than the true value due to the inclusion of zero estimates (results not shown). The post-selection inference is an active research area of its own and it is beyond the scope of this paper to identify the optimal post-selection inference procedure.

6 Analysis of ARIC Study

We use the proposed model selection procedures to analyze the ARIC study data (Ballantyne et al., 2004). As mentioned in the Introduction section, a cohort of 15,792 individuals were sampled from four U.S. communities and followed for ten years for the development of CHD. After excluding subjects for missing data and other reasons, a total of 12,199 subjects comprised the potential full cohort. Those who were alive and free of disease by the end of 1998 or lost to follow-up during the study periods were treated as censored. A random subcohort of 978 participants was selected by stratified random sampling from strata defined by sex, race (black versus white), and age at baseline (≤ 55 versus >55). After including all CHD cases, the case-cohort size is 1,568. There is a total of 638 CHD cases, corresponding to a censoring rate of 94.8%. In this analysis we are interested in identifying risk factors for incidence of CHD. In particular, the main risk factor of interest is the protein hs-CRP level, which is modeled as a categorical variable of low (<1.0 mg/L), middle (1.0 – 3.0 mg/L), and high (>3.0 mg/L) levels due to its nonlinear effect on the risk of CHD. Since CRP level is the main exposure variable, we do not penalize its regression coefficients and therefore set their tuning parameters to zero. Similarly, we keep the CRP terms in the model for the hard threshold method regardless of their p values. We also consider several other factors in the model selection process: age (years), BMI, systolic blood pressure (mmHg), LDL (mmol/L), HDL (mmol/L), diabetes (yes/no), and current smoker (yes/no). As shown in Kang et al. (2013), the empirical cumulative hazards functions for the different CRP groups increase approximately in a linear fashion. Therefore, the additive hazards model is a reasonable choice.

Table 5 summarizes the baseline characteristics of the full cohort and the subcohort. Note that the CRP level is not available for the full cohort due to the case-cohort design. It seems

that the distribution of the covariates are reasonably similar between the full cohort and subcohort, so the subcohort is representative of the full cohort.

We apply the SCAD penalized variable selection procedures with tuning parameter λ_n^{CV} or λ_n^{CVP} as well as the backwards elimination method to the ARIC study data to identify important risk factors for CHD. We include all covariates reported in Table 5 in the initial model. To ensure we do not miss any higher order effect of continuous variables and interactions between CRP and other variables, we include quadratic terms of all continuous variables as well as pairwise interaction between CRP and all other variables in the initial model. All continuous variables are standardized using the means and standard deviations based on the random subcohort. The tuning parameter selector identified $\lambda_n^{CV} = 0.993$ and $\lambda_n^{CVP} = 2.466$. Table 6 shows the selected covariates and their estimated coefficients and standard errors by the three methods. The SCAD with λ_n^{CV} selects the largest model and SCAD with λ_n^{CVP} selects the smallest model. This is consistent with the observation in the simulation study that the CV and backwards elimination methods tend to over-select variables compared to the CVP method. Besides CRP levels, all three methods identify HDL and HDL^2 as significant risk factors for CHD. The SCAD with λ_n^{CV} additionally includes 11 covariates and the backwards elimination method additionally selects 7 covariates.

Based on the results in Table 6 with λ_n^{CVP}, the risk of CHD for subjects whose serum CRP level is between 1.0 and 3.0 mg/L is 0.50×10^{-5} (95% CI: -0.99×10^{-5} to 1.99×10^{-5}) per-day, or 1.83 (95% CI: –3.61 to 7.27) per 1,000 person years, higher than those whose CRP level is below 1.0 mg/L. The risk of CHD for subjects whose CRP level is above 3.0 mg/L is 0.99×10^{-5} (95% CI: -0.50×10^{-5} to 2.48×10^{-5}) per-day, or 3.62 (95% CI: –1.82 to 9.06) per 1,000 person years, higher than those whose CRP level is below 1.0 mg/L. The effect of HDL level on risk of CHD follows a quadratic form with the minimum risk achieved at an HDL level of 2.7 standard deviations above population mean. Thus, vast majority of the population lie below this level. Hence there is a negative association between HDL level and risk of CHD, and the magnitude of the association decreases as HDL level increases. This finding is consistent with the common knowledge that HDL is the “good” cholesterol.

7 Discussion

In this paper we investigate a regularized variable selection procedure based on SCAD penalty in an additive hazards model with a stratified case-cohort design and a diverging number of parameters. Although this study is similar to the previous one on Cox proportional hazards model (Ni et al., 2016), it makes several important contributions to the field. First, as mentioned in the introduction, the additive hazards model is an important alternative to the proportional hazards model that warrants its own investigation, especially when the incidence rate is low as in most case-cohort studies. Second, due to the different objective functions used in the regularized estimation in the two models, there are differences in the theoretical derivations and tuning parameter selection methods. In

particular, the allowed divergence rate of the number of parameters to ensure the oracle property is higher in the additive hazards model than in the proportional hazards model. We also propose and assess a modified cross-validation tuning parameter selection method that is tailored to the additive hazards model with very high censoring rates. Lastly, in this paper we consider stratification by covariates observed in the entire cohort. This is an important strategy to increase efficiency of the case-cohort design, which was not investigated in the previous study on the Cox model.

In the simulation study we find that the proposed tuning parameter selection method outperforms the conventional cross-validation in terms of identifying the true model under all simulation scenarios. This is expected as the proposed modified cross-validation method incorporates a penalty on the model size to compensate for the overfitting effect of cross-validation. In public health and biomedical studies, investigators are often interested in identifying the true risk factors for the disease under study to facilitate policy making or reveal the underlying biological mechanisms of the disease. In such situations, it is more important to identify the true model than to predict the risk of individual subjects. Since we have demonstrated the superior performance of the penalized cross-validation method in model identification under various sample sizes, censoring rates, and sampling probabilities of the subcohort, we recommend it for practical use for tuning parameter selection in additive hazards model with a case-cohort design when the purpose of the study is to identify the true risk factors for the disease. A formal theoretical investigation on the proposed properties of the penalized cross-validation method is a future research topic.

In this paper we adopt the stratified case-cohort design proposed in Borgan et al. (2000). In that paper the authors focused on the estimation efficiency for a single “exposure” variable that is correlated with the stratification variable. However, in observational studies many covariates are likely to be correlated with the stratification variable, and therefore stratification considered in this paper is a general strategy to increase the estimation efficiency for all covariates in the model under a case-cohort design. The theoretical results in this paper require the number of strata to be finite and $n_s/n = O(1)$ for $s = 1, \ldots, S$, which implies that the condition $d_n^2/n_s \to 0$ is required for all s. The practical implication of this requirement is that the number of strata should not increase with sample size, which is a reasonable assumption since most stratification variables have a fixed number of pre-specified categories.

The proposed variable selection method does not have a mechanism to ensure the hierarchical structure of the candidate covariates in the presence of interaction and polynomial terms. As a result, the selected models from the ARIC study does not maintain a hierarchical structure. For example, the model identified by SCAD with λ_n^{CV} contains an interaction between CRP2 and BMI but not the main effect of BMI. Although this issue does not pose any theoretical difficulties, it does lead to some difficulties in the interpretation. A future research topic would be to take into account the hierarchical structure of the candidate covariates in an additive hazards model with a case-cohort design by using group variable selection techniques (Yuan & Lin, 2006; Zeng & Xie, 2014).

The R and Matlab programs implementing the estimation procedure described in Section 4.2 are available from the corresponding author upon request.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was partially supported by National Institutes of Health grants (P01 CA 142538, R01 ES 021900). The authors thank the staff and participants of the ARIC study for their important contributions. The ARIC Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (N01-HC-55015, N01-HC-55016, N01-HC-55018, N01-HC-55019, N01-HC-55020, N01-HC-55021, N01-HC-55022).

References

Aalen, O. Lecture Notes in Statistics 2. New York: Springer-Verlag; 1980. A model for nonparametric regression analysis of counting processes; p. 1-25.

Akaike H. Maximum likelihood identification of Gaussian autoregressive moving average models. Biometrika. 1973; 60:255–265.

Ballantyne CM, Hoogeveen RC, Bang H, Coresh J, Folsom AR, Heiss G, Sharrett AR. Lipoprotein-associated phospholipase a2, high-sensitivity c-reactive protein, and risk for incident coronary heart disease in middle-aged men and women in the Atherosclerosis Risk in Communities (ARIC) study. Circulation. 2004; 109:837–842. [PubMed: 14757686]

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

Cox DR. Regression models and life-tables. Journal of the Royal Statistical Society, Series B. 1972; 34:187–220.

Craven P, Wahba G. Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math. 1979; 31:377–403.

Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. Journal of the American Statistical Association. 2001; 96:1348–1360.

Hastie, T., Tibshirani, R., Friedman, J. The elements of statistical learning: data mining, inference and prediction. 2. Springer; 2009.

Huber PJ. Robust regression: Asymptotics, conjectures, and monte carlo. The Annals of Statistics. 1973; 1:799–821.

Kang S, Cai J, Chambless L. Marginal additive hazards model for case-cohort studies with multiple disease outcomes: an application to the atherosclerosis risk in communities (aric) study. Biostatistics. 2013; 14:28–41. [PubMed: 22826550]

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

Lin D, Ying Z. Semiparametric analysis of the additive risk model. Biometrika. 1994; 81:61–71.

Lin W, Lv J. High-dimensional sparse additive hazards regression. Journal of the American Statistical Association. 2013; 108:247–264.

Martinussen T, Scheike TH. Covariate selection for the semiparametric additive risk model. Scandinavian Journal of Statistics. 2009; 36:602–619.

Ni A, Cai J, Zeng D. Variable selection for case-cohort studies with failure time outcome. Biometrika. 2016; 103:547–562. [PubMed: 28529347]

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

Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society, Series B. 1996; 58:267–288.

Yuan M, Lin Y. Model selection and estimation in regression with grouped variables. Journal of the Royal Statistical Society, Series B. 2006; 68:49–67.

Zeng L, Xie J. Group variable selection via scad-l2. Statistics. 2014; 48:49–66.

Table 1

Model selection performance with $\beta_{\min} = 0.70$

Method	85% Censored RME median (MAD)	FP	FN	MS	RITM (%)	90% Censored RME median (MAD)	FP	FN	MS	RITM (%)
					$n = 10000$, $R = 1:1$					
	$\alpha = 0.18$, $d_n = 19$					$\alpha = 0.11$, $d_n = 16$				
CV	0.86 (0.16)	1.4	0.2	7	41.2	0.89 (0.17)	1.4	0.2	6	34.8
CVP	0.84 (0.23)	0.6	0.3	7	44.4	0.85 (0.24)	0.8	0.3	6	40.8
Back	0.76 (0.18)	0.6	0.1	7	54.4	0.74 (0.22)	0.5	0.2	6	52.0
Oracle	0.64 (0.19)	0.0	0.0	7	100.0	0.58 (0.2)	0.0	0.0	6	100.0
					$n = 10000$, $R = 1:2$					
	$\alpha = 0.35$, $d_n = 19$					$\alpha = 0.22$, $d_n = 16$				
CV	0.74 (0.18)	1.0	0.0	7	62.0	0.80 (0.2)	1.2	0.1	6	50.0
CVP	0.70 (0.22)	0.4	0.1	7	69.2	0.77 (0.23)	0.4	0.2	6	58.0
Back	0.71 (0.15)	0.6	0.0	7	53.6	0.73 (0.19)	0.5	0.0	6	56.8
Oracle	0.62 (0.19)	0.0	0.0	7	100.0	0.55 (0.24)	0.0	0.0	6	100.0
					$n = 15000$, $R = 1:1$					
	$\alpha = 0.18$, $d_n = 23$					$\alpha = 0.11$, $d_n = 19$				
CV	0.75 (0.18)	1.0	0.1	8	54.0	0.86 (0.19)	1.5	0.3	7	36.0
CVP	0.72 (0.19)	0.6	0.1	8	57.2	0.83 (0.30)	0.7	0.4	7	39.2
Back	0.72 (0.17)	0.8	0.0	9	43.2	0.74 (0.24)	0.6	0.2	7	48.4
Oracle	0.61 (0.21)	0.0	0.0	8	100.0	0.61 (0.23)	0.0	0.0	7	100.0
					$n = 15000$, $R = 1:2$					
	$\alpha = 0.35$, $d_n = 23$					$\alpha = 0.22$, $d_n = 19$				
CV	0.67 (0.21)	0.5	0.0	8	76.4	0.81 (0.18)	0.8	0.1	7	58.0
CVP	0.65 (0.19)	0.2	0.0	8	83.2	0.79 (0.22)	0.3	0.2	7	62.0
Back	0.70 (0.17)	0.8	0.0	9	45.2	0.75 (0.16)	0.5	0.0	7	55.2
Oracle	0.59 (0.20)	0.0	0.0	8	100.0	0.67 (0.19)	0.0	0.0	7	100.0

n: sample size; R: case to control ratio in the case-cohort sample; α: sampling proportion of random subcohort for both strata; d_n: number of parameters; RME: relative model error; MAD: median absolute deviation; FP: average number of zero parameters incorrectly identified as nonzero; FN: average number of nonzero parameters incorrectly identified as zero; MS: median model size; RITM: rate of

identifying true model; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection; Back: backwards elimination.

Table 2

Estimation result for $\beta_{\min} = 0.70$

Method	$\hat{\beta}_{\min}$	SE_e	SE_m	95% CI_e	$\bar{\lambda}_n$	$\hat{\beta}_{\min}$	SE_e	SE_m	95% CI_e	$\bar{\lambda}_n$
	85% Censored					90% Censored				
	$n = 10000$, $R = 1:1$									
	$\alpha = 0.18$, $d_n = 19$					$\alpha = 0.11$, $d_n = 16$				
CV	0.70	0.13	0.13	96.4	0.43	0.72	0.15	0.15	96.0	0.42
CVP	0.70	0.12	0.13	96.4	0.50	0.72	0.15	0.15	96.0	0.47
Back	0.70	0.12	0.13	97.2	–	0.72	0.15	0.15	96.0	–
Oracle	0.70	0.12	0.13	97.2	–	0.71	0.15	0.15	95.2	–
	$n = 10000$, $R = 1:2$									
	$\alpha = 0.35$, $d_n = 19$					$\alpha = 0.22$, $d_n = 16$				
CV	0.72	0.11	0.11	96.0	0.48	0.72	0.12	0.13	96.4	0.45
CVP	0.72	0.11	0.11	96.0	0.54	0.72	0.12	0.13	96.0	0.51
Back	0.72	0.11	0.11	95.6	–	0.72	0.12	0.13	97.2	–
Oracle	0.72	0.10	0.11	96.4	–	0.72	0.12	0.12	96.4	–
	$n = 15000$, $R = 1:1$									
	$\alpha = 0.18$, $d_n = 23$					$\alpha = 0.11$, $d_n = 19$				
CV	0.71	0.11	0.11	95.2	0.52	0.71	0.14	0.14	93.6	0.42
CVP	0.71	0.11	0.11	94.4	0.55	0.71	0.14	0.14	94.0	0.48
Back	0.71	0.11	0.11	93.6	–	0.71	0.14	0.14	93.2	–
Oracle	0.71	0.11	0.11	94.4	–	0.71	0.13	0.14	94.4	–
	$n = 15000$, $R = 1:2$									
	$\alpha = 0.35$, $d_n = 23$					$\alpha = 0.22$, $d_n = 19$				
CV	0.70	0.10	0.09	93.6	0.56	0.69	0.12	0.12	95.2	0.49
CVP	0.70	0.10	0.09	94.0	0.59	0.70	0.12	0.11	96.0	0.54
Back	0.70	0.10	0.09	94.4	–	0.70	0.12	0.12	95.6	–
Oracle	0.70	0.10	0.09	94.0	–	0.70	0.12	0.11	96.0	–

n: sample size; R: case to control ratio in the case-cohort sample; α: sampling proportion of random subcohort for both strata; d_n: number of parameters; SE_e: empirical standard error; SE_m: model-based standard error; 95% CI_e: empirical 95% coverage; $\bar{\lambda}_n$: average size of the tuning parameter; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection; Back: backwards elimination. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_{\min}$.

Table 3

Model selection performance with $\beta_{\min} = 0.43$

Method	85% Censored RME median (MAD)	FP	FN	MS	RITM (%)	90% Censored RME median (MAD)	FP	FN	MS	RITM (%)
					$n = 10000$, $R = 1:1$					
	$\alpha = 0.18$, $d_n = 19$					$\alpha = 0.11$, $d_n = 16$				
CV	0.89 (0.14)	1.9	0.3	8	28.8	0.87 (0.19)	2.0	0.4	8	24.4
CVP	0.87 (0.17)	0.9	0.4	7	34.4	0.85 (0.31)	0.8	0.6	8	27.6
Back	0.78 (0.18)	0.8	0.2	7	39.6	0.72 (0.22)	0.6	0.4	8	42.4
Oracle	0.65 (0.18)	0.0	0.0	7	100.0	0.53 (0.19)	0.0	0.0	8	100.0
					$n = 10000$, $R = 1:2$					
	$\alpha = 0.35$, $d_n = 19$					$\alpha = 0.22$, $d_n = 16$				
CV	0.78 (0.19)	1.2	0.1	7	51.2	0.81 (0.19)	1.3	0.2	8	45.2
CVP	0.76 (0.21)	0.4	0.2	7	58.0	0.76 (0.21)	0.5	0.3	8	51.2
Back	0.75 (0.16)	0.7	0.0	7	52.8	0.72 (0.19)	0.6	0.1	8	53.6
Oracle	0.62 (0.18)	0.0	0.0	7	100.0	0.6 (0.19)	0.0	0.0	8	100.0
					$n = 15000$, $R = 1:1$					
	$\alpha = 0.18$, $d_n = 23$					$\alpha = 0.11$, $d_n = 19$				
CV	0.72 (0.19)	1.0	0.2	8	53.2	0.89 (0.14)	1.9	0.3	6	30.0
CVP	0.70 (0.21)	0.5	0.2	8	54.8	0.86 (0.21)	0.9	0.5	6	31.2
Back	0.70 (0.17)	0.7	0.1	8	46.8	0.80 (0.21)	0.7	0.3	6	42.8
Oracle	0.59 (0.21)	0.0	0.0	8	100.0	0.64 (0.22)	0.0	0.0	6	100.0
					$n = 15000$, $R = 1:2$					
	$\alpha = 0.35$, $d_n = 23$					$\alpha = 0.22$, $d_n = 19$				
CV	0.67 (0.21)	0.8	0.0	8	70.8	0.86 (0.15)	1.3	0.2	6	44.8
CVP	0.63 (0.20)	0.3	0.0	8	77.6	0.79 (0.21)	0.6	0.3	6	50.4
Back	0.69 (0.17)	0.7	0.0	8	50.0	0.76 (0.18)	0.6	0.1	6	51.6
Oracle	0.60 (0.20)	0.0	0.0	8	100.0	0.66 (0.18)	0.0	0.0	6	100.0

n: sample size; R: case to control ratio in the case-cohort sample; α: sampling proportion of random subcohort for both strata; d_n: number of parameters; RME: relative model error; MAD: median absolute deviation; FP: average number of zero parameters incorrectly identified as nonzero; FN: average number of nonzero parameters incorrectly identified as zero; MS: median model size; RITM: rate of

identifying true model; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection; Back: backwards elimination.

Table 4

Estimation result for $\beta_{\min} = 0.43$

Method	85% Censored					90% Censored				
	$\hat{\beta}_{\min}$	SE_e	SE_m	95% CI_e	$\bar{\lambda}_n$	$\hat{\beta}_{\min}$	SE_e	SE_m	95% CI_e	$\bar{\lambda}_n$
	$n = 10000$, $R = 1:1$									
	$\alpha = 0.18$, $d_n = 19$					$\alpha = 0.11$, $d_n = 16$				
CV	0.47	0.12	0.13	95.9	0.40	0.49	0.14	0.15	94.6	0.38
CVP	0.48	0.11	0.13	95.7	0.47	0.51	0.13	0.15	93.4	0.47
Back	0.47	0.11	0.13	96.1	–	0.49	0.12	0.14	95.1	–
Oracle	0.44	0.13	0.13	93.2	–	0.44	0.15	0.14	93.6	–
	$n = 10000$, $R = 1:2$									
	$\alpha = 0.35$, $d_n = 19$					$\alpha = 0.22$, $d_n = 16$				
CV	0.44	0.11	0.11	95.7	0.47	0.47	0.11	0.12	95.6	0.43
CVP	0.45	0.11	0.11	95.5	0.53	0.48	0.10	0.12	97.2	0.50
Back	0.43	0.11	0.11	95.9	–	0.46	0.11	0.12	97.4	–
Oracle	0.42	0.11	0.11	94.8	–	0.45	0.12	0.12	94.8	–
	$n = 15000$, $R = 1:1$									
	$\alpha = 0.18$, $d_n = 23$					$\alpha = 0.11$, $d_n = 19$				
CV	0.44	0.10	0.11	97.8	0.50	0.47	0.13	0.14	92.6	0.40
CVP	0.44	0.09	0.10	97.7	0.54	0.48	0.12	0.14	93.4	0.47
Back	0.43	0.10	0.10	97.5	–	0.47	0.13	0.13	93.2	–
Oracle	0.42	0.10	0.10	96.0	–	0.44	0.14	0.13	93.2	–
	$n = 15000$, $R = 1:2$									
	$\alpha = 0.35$, $d_n = 23$					$\alpha = 0.22$, $d_n = 19$				
CV	0.43	0.09	0.09	96.7	0.54	0.45	0.11	0.11	96.5	0.45
CVP	0.43	0.08	0.09	97.1	0.57	0.45	0.10	0.11	96.4	0.50
Back	0.43	0.09	0.09	96.4	–	0.45	0.11	0.11	97.0	–
Oracle	0.43	0.08	0.09	96.8	–	0.43	0.11	0.11	95.2	–

n: sample size; R: case to control ratio in the case-cohort sample; α: sampling proportion of random subcohort for both strata; d_n: number of parameters; SE_e: empirical standard error; SE_m: model-based standard error; 95% CI_e: empirical 95% coverage; $\bar{\lambda}_n$: average size of the tuning parameter; CV: SCAD-penalized method with cross validation for tuning parameter selection; CVP: SCAD-penalized method with modified cross validation for tuning parameter selection; Back: backwards elimination. The parameter estimation results are calculated based on replications with nonzero $\hat{\beta}_{\min}$.

Table 5

Baseline characteristics of the cohort of ARIC study

Variables	Full cohort (n=12,199) Mean (SD) or %	Subcohort ($\tilde{n}$=978) Mean (SD) or %
Age (yrs)	56.8 (5.7)	58.4 (5.6)
BMI	27.9 (5.7)	28.2 (5.5)
Systolic blood pressure (mmHg)	132.8 (36.7)	134.6 (37.7)
LDL (mmol/L)	121.1 (18.5)	125.1 (19.8)
HDL (mmol/L)	50.5 (16.7)	49.5 (16.8)
Diabetes (%)	13.4	20.0
Current Smoker (%)	22.0	22.9
CRP level	–	3.3 (3.4)
CRP category (%)		
Low (<1.0mg/L)	–	28.2
Middle (1.0 – 3.0mg/L)	–	35.2
High (>3.0mg/L)	–	36.6

Table 6

Estimated coefficients and standard errors from ARIC study data

Variable	Backwards Elimination $\hat{\boldsymbol{\beta}}$ ($\widehat{\mathrm{SE}}$) (×10^{-5})	SCAD (λ_n^{CV}) $\hat{\boldsymbol{\beta}}$ ($\widehat{\mathrm{SE}}$) (×10^{-5})	SCAD (λ_n^{CVP}) $\hat{\boldsymbol{\beta}}$ ($\widehat{\mathrm{SE}}$) (×10^{-5})
CRP2 (middle (1.0 – 3.0mg/L))	−0.10 (0.21)	0.060 (0.82)	0.50 (0.76)
CRP3 (high (>3.0mg/L))	0.70 (0.24)	0.63 (0.81)	0.99 (0.76)
Age	0 (−)	0.27 (0.32)	0 (−)
Age^2	0.29 (0.11)	0 (−)	0 (−)
BMI	0 (−)	0 (−)	0 (−)
BMI^2	0 (−)	0 (−)	0 (−)
LDL (mmol/L)	0.72 (0.11)	0.43 (0.40)	0 (−)
LDL^2	0 (−)	0 (−)	0 (−)
HDL (mmol/L)	−1.42 (0.15)	−0.87 (0.63)	−1.27 (0.38)
HDL^2	0.34 (0.055)	0.22 (0.18)	0.24 (0.18)
SBP (mmHg)	0 (−)	0.46 (0.37)	0 (−)
SBP^2	0.34 (0.079)	0.14 (0.25)	0 (−)
Current Smoker	0.91 (0.25)	0 (−)	0 (−)
Diabetes	2.13 (0.39)	1.28 (0.85)	0 (−)
CRP2*age	0 (−)	0 (−)	0 (−)
CRP3*age	0 (−)	0 (−)	0 (−)
CRP2*BMI	−0.61 (0.18)	−0.56 (0.67)	0 (−)
CRP3*BMI	−0.35 (0.13)	−0.32 (0.43)	0 (−)
CRP2*LDL	0 (−)	0 (−)	0 (−)
CRP3*LDL	0 (−)	0.24 (0.66)	0 (−)
CRP2*HDL	0 (−)	−0.42 (0.76)	0 (−)
CRP3*HDL	0 (−)	−0.30 (0.79)	0 (−)
CRP2*SBP	0 (−)	0 (−)	0 (−)
CRP3*SBP	0 (−)	0 (−)	0 (−)
CRP2*current smoker	0 (−)	0.80 (1.32)	0 (−)
CRP3*current smoker	0 (−)	0 (−)	0 (−)
CRP2*diabetes	0 (−)	0 (−)	0 (−)
CRP3*diabetes	0 (−)	0 (−)	0 (−)

All continuous covariates are standardized using the means and standard deviations based on the random subcohort.