Published in final edited form as:

Biometrics. 2016 June ; 72(2): 372–381. doi:10.1111/biom.12452.

Kernel Machine Testing for Risk Prediction with Stratified Case Cohort Studies

Rebecca Payne, Matey Neykov, Majken Karoline Jensen, and Tianxi Cai*

Department of Biostatistics, Harvard School of Public Health, Boston, MA 02115

Summary

Large assembled cohorts with banked biospecimens offer valuable opportunities to identify novel markers for risk prediction. When the outcome of interest is rare, an effective strategy to conserve limited biological resources while maintaining reasonable statistical power is the case cohort (CCH) sampling design, in which expensive markers are measured on a subset of cases and controls. However, the CCH design introduces significant analytical complexity due to outcome-dependent, finite-population sampling. Current methods for analyzing CCH studies focus primarily on the estimation of simple survival models with linear effects; testing and estimation procedures that can efficiently capture complex non-linear marker effects for CCH data remain elusive. In this paper, we propose inverse probability weighted (IPW) variance component type tests for identifying important marker sets through a Cox proportional hazards kernel machine (Cox_{KM}) regression framework previously considered for full cohort studies (Cai et al., 2011). The optimal choice of kernel, while vitally important to attain high power, is typically unknown for a given dataset. Thus we also develop robust testing procedures that adaptively combine information from multiple kernels. The proposed IPW test statistics have complex null distributions that cannot easily be approximated explicitly. Furthermore, due to the correlation induced by CCH sampling, standard resampling methods such as the bootstrap fail to approximate the distribution correctly. We therefore propose a novel perturbation resampling scheme that can effectively recover the induced correlation structure. Results from extensive simulation studies suggest that the proposed IPW Cox_{KM} testing procedures work well in finite samples. The proposed methods are further illustrated by application to a Danish CCH study of Apolipoprotein C-III markers on the risk of coronary heart disease.

Keywords

Case Cohort; Cox Proportional Hazards Model; Inverse Probability Weighting; Finite Population Sampling; Kernel Machine Regression; Risk Prediction; Variance Component Test

* tcai@hsph.harvard.edu.

7. Supplementary Materials

Web Appendices referenced in Section 3 and the source code for the proposed testing procedures are available with this paper at the Biometrics website on Wiley Online Library.

1. Introduction

Achieving the full promise of personalized medicine requires tools for accurate risk assessment. When the risk of developing a clinical condition varies with the values of the risk markers, one could potentially use the marker information to develop optimal disease management strategies. Through collaborative efforts, several large prospective cohorts have been assembled over the past decades for studying risk markers of various diseases. Examples include the Atherosclerosis Risk in Communities (ARIC) study (Ballantyne et al., 2004), the Nurse's Health Study (NHS) (Colditz et al., 1997), and the Women's Health Initiative (WHI) (Anderson et al., 2003). These studies, with long term follow up information on a wide range of clinical conditions, have become a valuable resource for studying disease risks. In addition to recording phenotype information over time, biospecimens were often collected at baseline and stored for future studies. As knowledge accumulates over time, new markers may emerge as candidates for disease risk prediction. Once such research questions are formed, the stored specimens can be assayed and the marker measurements can be linked to phenotypes of interest to address these questions.

These large cohort studies provide valuable means for assessing the prognostic potential of novel markers. However, obtaining marker values for the entire cohort is often inefficient or infeasible due to cost constraints, especially if the clinical condition of interest is rare. For such settings, the two-phase case cohort (CCH) design is frequently employed as a cost-effective alternative to the full-cohort design. Examples of cohort studies that adopt the CCH design include the ARIC study for cardiovascular disease (Ballantyne et al., 2004) and the WHI for endometrial cancer (Gunter et al., 2008). To improve efficiency, the more general stratified CCH (sCCH) sampling is often adopted in practice to allow for sampling within strata defined by exposure variables and outcome status (Borgan et al., 2000; Samuelsen et al., 2007). Under the sCCH design, the clinical outcomes and routine clinical variables are typically collected for all subjects in the phase-I full cohort. However expensive markers are only measured for the phase-II subcohort, consisting of a fraction of cases and controls selected randomly within strata of the phase-I cohort. This design, while cost-effective, generates complex datasets for analysis due to outcome dependent stratified finite population sampling (IF-sampling). Here, the sCCH sampling is outcome dependent only through the event indicators and does not depend on the observed event times.

Regression methods for analyzing CCH data have been developed for the commonly used survival models (Prentice, 1986; Nan et al., 2009; Kong et al., 2004). Many of these methods employ inverse probability weighted (IPW) principles with weights accounting for the outcome dependent stratified sampling. General theory and applications of IPW estimators under sCCH design can be found in Breslow and Wellner (2007) and Gray (2009). Existing methods for analyzing CCH data typically assume that the markers have linear effects on a transformed survival time. When the markers affect the outcome non-linearly in a complex fashion, risk models and testing procedures derived under the linearity assumptions may have limited accuracy and power. One may incorporate non-linearity by including interactions or non-linear basis functions. However, it is often unclear which non-linear bases should be used for a given dataset and including a large number of non-informative bases could result in large estimation errors due to overfitting. In addition, when the sample

size of the phase-II cohort is not large, standard Wald test from fitting a model that includes a large number of non-linear basis functions may have incorrect size as shown in the simulation section. An effective approach to incorporating complex effects is through kernel machine (KM) regression (Schölkopf and Smola, 2001). The KM regression methods flexibly account for linear/non-linear effects, without requiring explicit specification of the basis functions. A wide range of KM regression methods have been developed for non-censored outcomes (Schölkopf and Smola, 2001). For survival outcomes, Cox proportional hazards KM (Cox_{KM}) regression and testing procedures have been studied for the standard full cohort setting (Cai et al., 2011; Li and Luan, 2003). However, no methods are currently available to analyze CCH data with KM models.

In this paper, we propose the use of the Cox_{KM} regression framework to derive a variance component score test for assessing the overall effect of a set of markers on survival with sCCH data. We overcome the outcome and strata dependent missingness in the markers via IPW and propose an IPW score statistic. One major challenge in carrying out the testing procedure is in deriving the null distribution of the score statistic. In general, making inference about IPW statistics under the sCCH design is difficult due to the weak correlation among the sampling indicators of different subjects induced by $\mathbb{F}$ -sampling. Ignoring such correlation corresponds to the use of the naive robust variance which could substantially over-estimate the true variance (Liu et al., 2012). As shown in the simulation studies presented in Section 4, the use of this robust variance could lead to substantially inflated type I error when testing for the overall effect of multiple markers. On the other hand, accounting for the correlations imposes a significant challenge to the estimation of the null distribution since explicit estimation is not feasible for the proposed score statistic, especially when non-linear kernels are employed. For such settings, resampling procedures such as the bootstrap would be helpful. However, the standard bootstrap is not directly applicable since it does not preserve the dependence structure of the data induced by the sampling.

For CCH studies, Wacholder et al. (1989) used a super-population model to construct a variance estimator for the log hazard ratio which is only valid in the absence of covariate effect. A modified bootstrap scheme was also proposed to approximate the distribution of the Prentice (1986) pseudo-likelihood estimator. However, no theoretical justification was provided and their approach cannot be easily extended to estimate the null distribution for the proposed score statistic. Recently, Huang (2014) proposed a bootstrap method for the original case cohort sampling design in which all cases and a random cohort are selected. However, their approach is not applicable to studies with sCCH sampling. To accommodate the general sCCH sampling, we propose a novel resampling procedure to estimate the null distribution of the IPW based score statistics. Our method is based on the equivalence between the IPW estimators constructed with true selection probabilities under sampling *without replacement* and the IPW estimators constructed with estimated sampling weights under sampling *with replacement* as detailed in Breslow and Wellner (2007). This motivates us to develop a procedure that mimics the effect of the correlation among sampling indicators by perturbing both the sampling indicator and the sampling probabilities.

The remainder of the paper is organized as follows. In Section 2, we describe the Cox_{KM} model and IPW estimation procedures for the model parameters. The variance component score statistic and the resampling procedures for approximating its null distribution are presented in Section 3. Adaptive methods for kernel tuning and selection to optimize power are also discussed. In Section 4, we present simulation results demonstrating that our proposed tests can maintain the desired type I error under the null and have good power in detecting both linear and non-linear effects. In Section 5 the proposed procedures are applied to a CCH study of apolipoprotein C-III markers for predicting the risk of CHD. Some concluding remarks are given in Section 6.

2. Cox_{KM} Modeling with CCH Data

2.1 Model Assumptions

Our primary goal is to examine whether a set of novel markers, $\mathbf{Z}_{p \times 1}$, is useful for predicting the event time T on top of a set of existing clinical variables $\mathbf{U}_{q \times 1}$ with sCCH data. To account for the potential non-linearity in the effect of $\mathbf{Z}$, we model the relationship between T and $\mathbf{W} = (\mathbf{U}^T, \mathbf{Z}^T)^T$ through a Cox_{KM} regression model (Li and Luan, 2003; Cai et al., 2011):

$$\lambda_{\mathbf{W}}(t) = \lambda_0(t) e^{\gamma_0^T \mathbf{U} + h_0(\mathbf{Z})}, \quad (1)$$

where $\lambda_{\mathbf{W}}(t)$ is the conditional hazard function of T given $\mathbf{W}$, $\lambda_0(\cdot)$ is an unknown baseline hazard function, and γ_0 is an unknown log-hazard ratio parameter representing the effect of $\mathbf{U}$ on T . Although the effect of the original set of clinical predictors may not always be linear, we assume that their effect on the outcome is sufficiently well understood and we use $\mathbf{U}$ to represent properly transformed clinical variables to ensure linearity. Here, the unknown function $h_0(\mathbf{Z})$, summarizing the effect of $\mathbf{Z}$ on T , is a centered smooth function belonging to a Reproducible Kernel Hilbert Space (RKHS) $\mathcal{H}_K$ generated by a given positive definite kernel function $K(\cdot, \cdot; \rho)$, and ρ is some tuning parameter (Cristianini and Shawe-Taylor, 2000). The kernel function $K(\mathbf{Z}_1, \mathbf{Z}_2; \rho)$ measures the similarity between $\mathbf{Z}_1$ and $\mathbf{Z}_2$ and different choices of K lead to different RKHS. Some of the popular kernel functions include the Gaussian kernel $K(\mathbf{Z}_1, \mathbf{Z}_2; \rho) = \exp\{-0.5(\mathbf{Z}_1 - \mathbf{Z}_2)^T(\mathbf{Z}_1 - \mathbf{Z}_2)/\rho\}$, which can be used to capture complex smooth non-linear effects; the linear kernel $K(\mathbf{Z}_1, \mathbf{Z}_2; \rho) = \rho + \mathbf{Z}_1^T \mathbf{Z}_2$ which corresponds to $h_0(\mathbf{Z})$ being linear in $\mathbf{Z}$; and the quadratic kernel $K(\mathbf{Z}_1, \mathbf{Z}_2; \rho) = (\mathbf{Z}_1^T \mathbf{Z}_2 + \rho)^2$, which allows for 2-way interactive effects. From here onward, for notational ease, unless noted otherwise we suppress ρ from the kernel function K .

By Mercer's Theorem (Cristianini and Shawe-Taylor, 2000), any $h(\mathbf{Z}) \in \mathcal{H}_K$ has a *primal representation* with respect to the eigensystem of K . Specifically, under the probability measure of $\mathbf{Z}$, K has eigenvalues $\{\lambda_l\}_{l=1, \dots, \mathcal{L}}$ with $\lambda_1 \geq \lambda_2 \geq \dots \lambda_{\mathcal{L}}$ and the corresponding eigenfunctions $\{\phi_l\}_{l=1, \dots, \mathcal{L}}$ such that $K(\mathbf{Z}_1, \mathbf{Z}_2) = \sum_{l=1}^{\mathcal{L}} \lambda_l \phi_l(\mathbf{Z}_1) \phi_l(\mathbf{Z}_2)$, where $\mathcal{L} \leq \infty$.

and $\lambda_l > 0$ for any $l < \infty$. The basis functions, $\{\psi_l(\cdot) = \sqrt{\lambda_l} \phi_l(\cdot)\}_{l=1, \dots, \mathcal{L}}$, span $\mathcal{H}_K$. Hence, model (1) is equivalent to

$$\lambda_{\mathbf{w}}(t) = \lambda_0(t) e^{\gamma_0^T \mathbf{U} + h_0(\mathbf{Z}_i)} = \lambda_0(t) e^{\gamma_0^T \mathbf{U} + \sum_{l=1}^{\mathcal{L}} \beta_{0l} \psi_l(\mathbf{Z}_i)}. \quad (2)$$

Throughout, we assume that $\mathbf{Z}$ is bounded and K is twice continuously differentiable, leading to bounded $\{\psi_l(\mathbf{Z})\}$ on the support of $\mathbf{Z}$.

2.2 Data Structure and Case Cohort Design

We assume that there are a total of N subjects in the phase I full cohort. Due to right censoring, the event time T is only observable up to a bivariate vector (X, δ) , where $X = T \wedge C$, $\delta = I(T < C)$ and C is the censoring time. The underlying full cohort data consists of N independent and identically distributed (i.i.d) random vectors,

$\mathcal{D} = \{\mathbf{D}_i = (X_i, \delta_i, \mathcal{J}_i, \mathbf{W}_i = (\mathbf{U}_i^T, \mathbf{Z}_i^T)^T)^T\}_{i=1, \dots, N}$, where $\mathcal{J}$ is a stratification variable used for CCH sampling that takes J unique values, $1, \dots, J$. Without loss of generality, we focus on the setting where $\{\mathbf{U}, X, \delta\}$ is collected for all subjects in the cohort but $\mathbf{Z}$ is only measured for subjects selected into the CCH subcohort. For $i = 1, \dots, N$, let V_i be a binary variable indicating whether the i th subject is selected into the CCH subcohort. Thus, the observable data consist of $\mathcal{D}_{CCH} = \{(X_i, \delta_i, \mathcal{J}_i, \mathbf{U}_i^T, V_i \mathbf{Z}_i^T, V_i)^T\}_{i=1, \dots, N}$ and without loss of generality we let $\Omega_{CCH} = \{1, \dots, n\} \subset \{1, 2, \dots, N\}$ be the index set for all subjects belonging to the sCCH subcohort. Note that when $V_i = 0$, $V_i \mathbf{Z}_i^T = 0$ and the value of $\mathbf{Z}_i$ is not observed.

We consider the general sCCH sampling scheme where the sampling is performed conditional on both δ and the stratification variable $\mathcal{J}$. For $d = 0, 1$ and $j = 1, \dots, J$, let

$\mathcal{J}_{dj,i} = I(\mathcal{J}_i = j, \delta_i = d)$, where $I(\cdot)$ is the indicator function. Let $N_{dj} = \sum_{i=1}^N \mathcal{J}_{dj,i}$ with N_{1j} and N_{0j} representing the number of cases and controls in the j th stratum, respectively. Within each stratum defined by $\mathcal{J} = j$, suppose n_{1j} out of N_{1j} cases and n_{0j} out of N_{0j}

controls were sampled *without replacement* into CCH subcohort. Let $n_d = \sum_{j=1}^J n_{dj}$ and $n = n_0 + n_1$. Conditional on the full data $\mathcal{D}$,

$$P(V_i = 1 | \mathcal{D}) = \hat{p}_i = \sum_{d=0}^1 \sum_{j=1}^J \mathcal{J}_{dj,i} \hat{\pi}_{dj},$$

where $\hat{\pi}_{dj} = n_{dj} / N_{dj}$. Let $\hat{\omega}_i = V_i / \hat{p}_i$ denote the weight used for the IPW estimators. Due to $\mathbb{F}$ -sampling, V_i and V_j are negatively correlated when $\mathcal{J}_{dj,i} = \mathcal{J}_{dj,j}$ with covariance

$$\text{Cov}(V_i, V_j | \mathcal{D}) = - \sum_{d,j} \hat{\pi}_{dj} \left(\hat{\pi}_{dj} - \frac{n_{dj} - 1}{N_{dj} - 1} \right) \mathcal{J}_{dj,i} \mathcal{J}_{dj,j} = - \sum_{d,j} N_d^{-1} \hat{\pi}_{dj} (1 - \hat{\pi}_{dj}) \mathcal{J}_{dj,i} \mathcal{J}_{dj,j},$$

which needs to be accounted for when deriving inference procedures (Liu et al., 2012).

Estimation of Model Parameters

To account for sCCH sampling, we propose to make inference about the model parameters based on a penalized weighted log partial likelihood (WLPL)

$$\sum_{i \in \Omega_{CCH}} \hat{\omega}_i \int \log \left[\frac{\exp\{h(\mathbf{Z}_i) + \gamma^\top \mathbf{U}_i\}}{\sum_{j \in \Omega_{CCH}} \hat{\omega}_j Y_j(s) \exp\{h(\mathbf{Z}_j) + \gamma^\top \mathbf{U}_j\}} \right] dN_i(s) - \frac{\vartheta}{2} \|h\|_{\mathcal{H}_K}^2, \quad \vartheta \geq 0, \quad (3)$$

where $Y_j(t) = I(X_j \geq t)$, $N_j(t) = \delta_j I(X_j \leq t)$, $\|h\|_{\mathcal{H}_K}$ is the RKHS norm of h , and the tuning parameter ϑ controls the amount of penalty for the smoothness in h . Since the basis functions $\{\psi_l\}$ depend on the unknown distribution of $\mathbf{Z}$, direct maximization of (3) based on the primal representation of h is infeasible. On the other hand, by the representer theorem (Kimeldorf and Wahba, 1970), we may show that the maximizer in (3) takes the dual representation with $h(\mathbf{Z}_i) = \mathbf{K}_i^\top \boldsymbol{\alpha}$, where $\mathbf{K}_i = (K_{i1}, \dots, K_{in})^\top$, $K_{ij} = K(\mathbf{Z}_i, \mathbf{Z}_j)$, and $\boldsymbol{\alpha}_{n \times 1}$ is the vector of unknown coefficients to be estimated. Note that K_{ij} is only defined when $\{i, j\} \subset \Omega_{CCH}$. It follows that maximizing (3) can be equivalently solved as an explicit optimization problem in the dual form:

$$\sum_{i \in \Omega_{CCH}} \hat{\omega}_i \int \log \left[\frac{\exp(\mathbf{K}_i^\top \boldsymbol{\alpha} + \gamma^\top \mathbf{U}_i)}{\sum_{j \in \Omega_{CCH}} \hat{\omega}_j Y_j(s) \exp(\mathbf{K}_j^\top \boldsymbol{\alpha} + \gamma^\top \mathbf{U}_j)} \right] dN_i(s) - \frac{\vartheta}{2} \boldsymbol{\alpha}^\top \mathbb{K} \boldsymbol{\alpha}, \quad (4)$$

where $\mathbb{K} = [\mathbf{K}_1, \dots, \mathbf{K}_n]_{n \times n}$. Let $\hat{\boldsymbol{\alpha}}$ and $\hat{\gamma}$ be the maximizers of (4) for a given ϑ . Then for any $\mathbf{Z}$, $h(\mathbf{Z})$ can be estimated as $\hat{h}(\mathbf{Z}) = \sum_{i=1}^n K(\mathbf{Z}_i, \mathbf{Z}) \hat{\alpha}_i$. The log-likelihood in (4) can be also viewed as a penalized quasi-likelihood approximation to a WLPL corresponding to the random effects model with $\boldsymbol{\alpha} \sim \mathcal{N}(\mathbf{0}, \vartheta^{-1} \mathbb{K}^-)$. The tuning parameter ϑ can be selected based on standard methods such as the Bayesian information criterion (BIC), Akaike information criterion (AIC) or the cross-validation procedures. If all subjects are sampled into the CCH subcohort, these estimators reduce to the Li and Luan (2003) and Cai et al. (2011) estimators under the full cohort design.

The dual estimator can also be viewed as the solution to the primal problem with $\{\psi_l(\cdot)\}$ estimated via a weighted kernel spectral decomposition with weights accounting for the sampling. To see this, let $\{\hat{\lambda}_l, \hat{\phi}_l = (\hat{\phi}_l(\mathbf{Z}_1), \dots, \hat{\phi}_l(\mathbf{Z}_n))^\top\}_{l=1, \dots, n}$ be the eigenvalue and eigenvectors corresponding to $\mathbb{K}_w = N^{-1} \mathbb{K} \mathbb{D}_w$ with $\mathbb{K}_w \hat{\phi}_l = \hat{\lambda}_l \hat{\phi}_l$, where $\mathbb{D}_w = \text{diag}\{\hat{\omega}_1, \dots, \hat{\omega}_n\}$. Since $E\{\hat{\omega}_i | \mathcal{D}\} = 1$, $\hat{\psi}_l = \hat{\lambda}_l^{1/2} \hat{\phi}_l$ is an estimator for $\psi_l = (\psi_l(\mathbf{Z}_1), \dots, \psi_l(\mathbf{Z}_n))^\top$. It then follows a variable transformation, $\hat{\Psi} \boldsymbol{\beta} = \mathbb{K} \boldsymbol{\alpha}$, that (4) can be equivalently written as

$$\sum_{i \in \Omega_{CCH}} \hat{\omega}_i \int \log \left[\frac{\exp(\hat{\Psi}_i^\top \beta + \gamma^\top \mathbf{U}_i)}{\sum_{j \in \Omega_{CCH}} \hat{\omega}_j Y_j(s) \exp(\hat{\Psi}_j^\top \beta + \gamma^\top \mathbf{U}_j)} \right] dN_i(s) - \frac{\vartheta}{2} \beta^\top \beta. \quad (5)$$

where $\hat{\Psi} = [\hat{\psi}_1, \dots, \hat{\psi}_n]$ and $\hat{\Psi}_i = (\hat{\psi}_1(\mathbf{Z}_i), \dots, \hat{\psi}_n(\mathbf{Z}_i))^\top$ is the i th row vector of $\hat{\Psi}$. Let $\hat{\beta}$ and $\hat{\gamma}$ be the maximizer of (5). Then $\hat{\beta}$ can be viewed as a ridge penalized partial likelihood estimate for the coefficient vector $\beta_0^{[n]} = (\beta_{01}, \dots, \beta_{0n})$. Subsequently, we estimate $h(\mathbf{Z})$ as $\hat{h}(\mathbf{Z}) = \hat{\beta}^\top \hat{\Psi}(\mathbf{Z})$, which can be shown as equivalent to $\hat{h}(\mathbf{Z}) = \sum_{i=1}^n K(\mathbf{Z}_i, \mathbf{Z}) \hat{\alpha}_i$ obtained under dual representation.

Since the eigenvalues may decay quickly, $\mathcal{H}_K$ can be approximated well by the span of $\{\psi_l, l = 1, \dots, r\}$ for some $r \ll n$. Thus, we may estimate $\mathbf{h} = (h_0(\mathbf{Z}_1), \dots, h_0(\mathbf{Z}_n))^\top$ more efficiently as $\hat{\mathbf{h}} = \hat{\Psi}^{[r_n]} \hat{\beta}^{[r_n]}$ by performing a kernel principal component analysis (PCA) (Braun, 2005), where $\hat{\Psi}^{[r_n]} = [\hat{\psi}_1, \dots, \hat{\psi}_{r_n}]$, $\hat{\beta}^{[r_n]}$ is the minimizer of

$$\sum_{i \in \Omega_{CCH}} \hat{\omega}_i \int \log \left[\frac{\exp(\beta^\top \hat{\Psi}_i^{[r_n]} + \gamma^\top \mathbf{U}_i)}{\sum_{j \in \Omega_{CCH}} \hat{\omega}_j Y_j(s) \exp(\beta^\top \hat{\Psi}_j^{[r_n]} + \gamma^\top \mathbf{U}_j)} \right] dN_i(s) - \frac{\vartheta}{2} \beta^\top \beta. \quad (6)$$

and $\hat{\Psi}_i^{[r_n]} = (\hat{\psi}_1(\mathbf{Z}_i), \dots, \hat{\psi}_{r_n}(\mathbf{Z}_i))^\top$. To estimate $h_0(\mathbf{Z})$ for a future $\mathbf{Z}$, we use the Nystrom method (Rasmussen and Williams, 2006) along IPW and obtain $\hat{h}_0(\mathbf{Z}) = \hat{\psi}^{[r_n]}(\mathbf{Z}) \hat{\beta}^{[r_n]}$, where $\hat{\Psi}^{[r_n]}(\mathbf{Z})^\top = N^{-1} [\hat{\lambda}_1^{-1} \hat{\psi}_1, \dots, \hat{\lambda}_{r_n}^{-1} \hat{\psi}_{r_n}]^\top \mathbb{D}_w[k(\mathbf{Z}, \mathbf{Z}_1), \dots, k(\mathbf{Z}, \mathbf{Z}_n)]^\top$. In practice, r_n may be selected as the smallest r such that $\sum_{l=1}^r \hat{\lambda}_l / \{\sum_{l=1}^n \hat{\lambda}_l\} \geq \varphi$ for some pre-determined $\varphi \in (0, 1)$.

3. Cox_{KM} Score Test with CCH Data

3.1 Score Statistic

To derive a test for $H_0 : h_0(\cdot) = 0$, we propose a weighted variance component score test by imposing a *working* assumption that $\alpha = \tau \varepsilon$ and $\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^\top$ is a random vector with mean 0 and variance covariance matrix $\mathbb{K}^-$. Under this working assumption, H_0 is equivalent to

$$H_0^\dagger : \tau = 0.$$

The WLPL function conditional on ε can be written as

$$\mathcal{L}_\varepsilon(\tau; \gamma) = \sum_{i \in \Omega_{CCH}} \hat{\omega}_i \int \log \left[\frac{\exp(\tau \mathbf{K}_i^\top \varepsilon + \gamma^\top \mathbf{U}_i)}{\sum_{j \in \Omega_{CCH}} \hat{\omega}_j Y_j(s) \exp(\tau \mathbf{K}_j^\top \varepsilon + \gamma^\top \mathbf{U}_j)} \right] dN_i(s). \quad (7)$$

For a given γ , a normalized score statistic for τ can be obtained as

$$\hat{\mathcal{Q}}_w(\gamma) = N^{-1} E \left\{ \dot{\mathcal{L}}_\varepsilon(0; \gamma)^2 | \mathcal{D}_{CCH} \right\}, \quad \text{where} \quad \dot{\mathcal{L}}_\varepsilon(\tau; \gamma) = \partial \mathcal{L}_\varepsilon(\tau; \gamma) / \partial \tau.$$

Here, the expectation is taken over the distribution of ε . It is not difficult to show that

$$\hat{\mathcal{Q}}_w(\gamma) = N^{-1} \left\{ \mathbb{D}_w \hat{\mathbf{M}}_w(\infty; \gamma) \right\}^\top \mathbb{K} \left\{ \mathbb{D}_w \hat{\mathbf{M}}_w(\infty; \gamma) \right\}$$

where $\hat{\mathbf{M}}_w(t; \gamma) = [\hat{M}_{1w}(t; \gamma), \dots, \hat{M}_{nw}(t; \gamma)]^\top$,
 $\hat{M}_{wi}(t; \gamma) = N_i(t) - \int_0^t e^{\gamma^\top \mathbf{U}_i} Y_i(t) d\hat{\Lambda}_{0w}(t; \gamma)$, $\mathbb{D}_w = \text{diag}\{\hat{\omega}_1, \dots, \hat{\omega}_n\}$ and
 $\hat{\Lambda}_{0w}(t; \gamma) = \int_0^t \sum_{i=1}^N \hat{\omega}_i dN_i(s) / \left\{ \sum_{j=1}^N \hat{\omega}_j Y_j(s) e^{\gamma^\top \mathbf{U}_j} \right\}$. Since γ_0 is often unknown, may estimate it by maximizing the WLPL under the null and then replace γ in $\hat{\mathcal{Q}}_w(\gamma)$ with this estimator, denoted by $\hat{\gamma}_w$, to obtain the score statistic $\hat{\mathcal{Q}}_w = \hat{\mathcal{Q}}_w(\hat{\gamma}_w)$.

The estimates $\hat{\gamma}_w$ and $\hat{\Lambda}_{0w}(t; \gamma)$ in $\hat{\mathcal{Q}}_w$ are constructed using only observations in the CCH subcohort. When $\{\mathbf{U}, X, \delta\}$ is observable for the full cohort, we may improve the efficiency by using the full cohort data to estimate γ with the standard partial likelihood estimator, $\hat{\gamma}$, and $\hat{\Lambda}_0(t)$ as the Breslow estimator, $\hat{\Lambda}_0(t) = \hat{\Lambda}_0(t; \hat{\gamma})$ where

$\hat{\Lambda}_0(t; \gamma) = \sum_{i=1}^N \int_0^t \left\{ \sum_{j=1}^n Y_j(s) e^{\gamma^\top \mathbf{U}_j} \right\}^{-1} dN_i(s)$. Let the resulting martingale residual be denoted as $\hat{\mathbf{M}} = \hat{\mathbf{M}}(\infty; \hat{\gamma})$, $\hat{\mathbf{M}}(t; \gamma) = [\hat{M}_1(t; \gamma), \dots, \hat{M}_n(t; \gamma)]^\top$ and
 $\hat{M}_i(t; \gamma) = N_i(t) - \int_0^t e^{\gamma^\top \mathbf{U}_i} Y_i(t) d\hat{\Lambda}_0(t; \gamma)$. Replacing $\hat{\gamma}_w$ and $\hat{\Lambda}_{0w}(\cdot)$ in $\hat{\mathcal{Q}}_w$ with $\hat{\gamma}$ and $\hat{\Lambda}_0(t)$ respectively, we propose our final score test statistic as

$$\hat{\mathcal{Q}} = N^{-1} \left(\mathbb{D}_w \hat{\mathbf{M}} \right)^\top \mathbb{K} \left(\mathbb{D}_w \hat{\mathbf{M}} \right). \quad (8)$$

3.2 Approximating the Null distribution

To approximate the null distribution of $\hat{\mathcal{Q}}$, we show in Web Appendix A that under H_0 ,

$$\hat{\mathcal{Q}} = \int \int K(\mathbf{Z}_1, \mathbf{Z}_2) d\hat{\mathbf{C}}(\mathbf{Z}_1) d\hat{\mathbf{C}}(\mathbf{Z}_2) + o_p(1),$$

where $\hat{\mathbb{C}}(\cdot)$ converges weakly to a zero-mean Gaussian process, denoted by $\mathcal{C}(\cdot)$. Thus the null distribution of $\hat{\mathcal{Q}}$ can be approximated by the distribution of

$\int \int K(\mathbf{Z}_1, \mathbf{Z}_2) d\mathcal{C}(\mathbf{Z}_1) d\mathcal{C}(\mathbf{Z}_2)$. However, this distribution is typically difficult to estimate explicitly as it involves integrations of the variance covariance functions of $\mathcal{C}(\cdot)$.

Furthermore, due to correlation among the sampling indicators $\{V_i\}_{i=1, \dots, N}$, standard resampling procedures such as the bootstrap lead to inconsistent estimation of variances. Motivated by the equivalence results given in Breslow and Wellner (2007), we propose a novel resampling procedure that recovers the effect of the correlation among $\{V_i\}_{i=1, \dots, N}$.

To provide some intuition for our proposal, consider a simple statistic $\hat{\theta} = N^{-1} \sum_{i=1}^n \hat{\omega}_i R_i$, where $R_i = R(\mathbf{W}_i, X_i, \delta_i)$ for some function R that has total variation bounded and $E(R_i) = 0$. Furthermore, consider two sampling schemes to generate $\{V_i\}$: (I) $\mathbb{F}$ -sampling; and (II) Bernoulli sampling ($\mathbb{B}$ -sampling) with $V_i \sim \text{independent Bernoulli}(\hat{p}_i)$. From Breslow and Wellner (2007), $N^{\frac{1}{2}} \hat{\theta} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_R^2 - \sigma_C^2)$ under $\mathbb{F}$ -sampling and $N^{\frac{1}{2}} \hat{\theta} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma_R^2)$ under $\mathbb{B}$ -sampling, where σ_R^2 is the robust variance ignoring the negative correlation among $\{V_i\}$ and σ_C^2 accounts for the correlation. Under $\mathbb{F}$ -sampling, the naive bootstrap variance estimator approximates σ_R^2 . In addition, Breslow and Wellner (2007) showed that under

$\mathbb{B}$ -sampling, if we replace $\hat{\omega}_i$ by an estimated weight $\tilde{w}_i = V_i / \left\{ \sum_{d,j} \mathcal{J}_{dj,i} \tilde{\pi}_{dj} \right\}$ in $\hat{\theta}$, then the corresponding estimator has the same limiting distribution as $\hat{\theta}$ under $\mathbb{F}$ -sampling, where $\tilde{\pi}_{dj} = \left\{ \sum_{i=1}^n V_i \mathcal{J}_{dj,i} \right\} / \left\{ \sum_{i=1}^n \mathcal{J}_{dj,i} \right\}$ is the estimated sampling probability. This motivates us to use independent perturbations to the observations but recover the effect of the correlation due to $\mathbb{F}$ -sampling by estimating the sampling probabilities using the perturbed observations. We next detail the resampling procedure for approximating the distribution of $\hat{\mathcal{Q}}$.

Let $\mathbf{B} = (B_1, \dots, B_N)^T$ be a vector of i.i.d. non-negative random variables with unit mean and variance generated independently of the data. We obtain the perturbed counterpart of $\hat{\mathcal{Q}}$ as

$$\hat{\mathcal{Q}}^* = N^{-1} \left(\mathbb{D}_w^* \hat{\mathbf{M}}^* - \mathbb{D}_w \hat{\mathbf{M}} \right)^T \mathbb{K} \left(\mathbb{D}_w^* \hat{\mathbf{M}}^* - \mathbb{D}_w \hat{\mathbf{M}} \right)$$

where $\mathbb{D}_w^* = \text{diag}(\hat{\omega}_1^*, \dots, \hat{\omega}_n^*)$, $\hat{\mathbf{M}}^* = (\hat{M}_1^*, \dots, \hat{M}_n^*)^T$,

$$\hat{M}_i^* = \hat{M}_i - N^{-\frac{1}{2}} \int Y_i(t) e^{\gamma U_i} d\hat{W}_{\hat{\Lambda}}^*(t) - \hat{\mathcal{Q}}_i^T N^{-\frac{1}{2}} \hat{W}_{\gamma}^*, \quad \hat{w}_i^* = V_i B_i / \hat{p}_i^*, \quad \hat{p}_i^* = \sum_{dj} \mathcal{J}_{dj,i} \hat{\pi}_{dj}^*,$$

$$\hat{\pi}_{dj}^* = \sum_{i=1}^N B_i V_i \mathcal{J}_{dj,i} / \left\{ \sum_{j=1}^N B_j \mathcal{J}_{dj,j} \right\},$$

$$\begin{aligned}\hat{\mathbf{W}}_{\gamma}^* &= N^{-\frac{1}{2}} \sum_{j=1}^N \hat{\mathbf{W}}_{\gamma j} (B_j - 1), & \hat{\mathbf{W}}_{\Lambda}^* (t) &= N^{-\frac{1}{2}} \sum_{j=1}^N \hat{W}_{\Lambda j} (t) (B_j - 1) \\ \hat{W}_{\gamma j} &= \hat{\mathbf{A}}^{-1} \int \left\{ \mathbf{U}_j - \frac{\hat{S}^{(1)}(t)}{\hat{S}^{(0)}(t)} \right\} d\hat{M}_j(t), & \hat{W}_{\Lambda j} (t) &= \int_0^t \frac{d\hat{M}_j(u) - \hat{W}_{\gamma j} \hat{S}^{(1)}(u) d\hat{\Lambda}_0(u)}{\hat{S}^{(0)}(u)} \\ \hat{\mathbf{A}} &= \int \frac{\hat{S}^{(2)}(u) \hat{S}^{(0)}(u) - \hat{S}^{(1)}(u)^{\otimes 2}}{\hat{S}^{(0)}(u)^2} d\bar{N}(u), & \hat{\mathcal{Q}}_i &= \mathbf{U}_i \int Y_i(t) e^{\hat{\gamma} \mathbf{U}_i} d\hat{\Lambda}_0(t),\end{aligned}$$

$\hat{M}_i(t) = \hat{M}_i(t, \hat{\gamma})$, $\hat{S}^{(k)}(u) = \hat{S}^{(k)}(u; \hat{\gamma})$, $\hat{S}^{(k)}(u, \gamma) = N^{-1} \sum_{i=1}^N I(X_i \geq u) e^{\gamma^\top \mathbf{U}_i} \mathbf{U}_i^{\otimes k}$, and for any vector $\mathbf{a}$, $\mathbf{a}^{\otimes 0} = 1$, $\mathbf{a}^{\otimes 1} = \mathbf{a}$ and $\mathbf{a}^{\otimes 2} = \mathbf{a}\mathbf{a}^\top$. We show in Web Appendix B that distribution of $\hat{\mathcal{Q}}^*$ conditional on $\mathcal{D}_{CCH}$ can be used to approximate the null distribution of $\hat{\mathcal{Q}}$. In practice, one may generate $\mathcal{B}$ realizations of $\mathbf{B}$, and calculate $\mathcal{B}$ instances of $\hat{\mathcal{Q}}^*$,

denoted by $\{\hat{\mathcal{Q}}^{*b}\}_{b=1, \dots, \mathcal{B}}$. The p -value of the test can be calculated as

$\mathcal{B}^{-1} \sum_{b=1}^{\mathcal{B}} I(\hat{\mathcal{Q}}^{*b} > \hat{\mathcal{Q}})$. In finite samples, we found that the null distribution can be better approximated by a slight modification of $\hat{\mathcal{Q}}^*$. We may write $\hat{\mathcal{Q}} = \hat{\mathcal{Q}}_1 + \hat{\mathcal{Q}}_2$ and show that

$\hat{\mathcal{Q}}_1 \rightarrow q_1$ in probability, where $\hat{\mathcal{Q}}_1 = N^{-1} \sum_{i=1}^n \hat{M}_i^2 K(\mathbf{Z}_i, \mathbf{Z}_i) \hat{\omega}_i^2$,
 $\hat{\mathcal{Q}}_2 = N^{-1} \sum_{i \neq j} \hat{M}_i \hat{M}_j K(\mathbf{Z}_i, \mathbf{Z}_j) \hat{\omega}_i \hat{\omega}_j$, $q_1 = E\{M_i^2 K(\mathbf{Z}_i, \mathbf{Z}_i) / p_i\}$, $M_i = \delta_i - \Lambda_0(X_i) e^{\gamma_0^\top \mathbf{U}_i}$
and p_i is limit of $\hat{p}_i$. Since $q_1 + \hat{\mathcal{Q}}_2$ and $\hat{\mathcal{Q}}$ have the same limiting distribution, we propose to instead obtain $\hat{\mathcal{Q}}^* = \hat{\mathcal{Q}}_1 + \hat{\mathcal{Q}}_2^*$ to approximate the distribution of $\hat{\mathcal{Q}}$, where
 $\hat{\mathcal{Q}}_2^* = N^{-1} \sum_{i \neq j} \hat{M}_i \hat{M}_j K(\mathbf{Z}_i, \mathbf{Z}_j) \hat{\omega}_i^* \hat{\omega}_j^*$.

3.3 Tuning and Kernel Selection

For kernels that involve a tuning parameter ρ , let $\hat{\mathcal{Q}}_\rho$ denote the score statistic obtained using the kernel matrix $\mathbb{K}(\rho) = [K(\mathbf{Z}_i, \mathbf{Z}_j; \rho)]$ and $\{\hat{\mathcal{Q}}_\rho^{*b}\}_{b=1, \dots, \mathcal{B}}$ denote the $\mathcal{B}$ realizations of the perturbed counterpart of $\hat{\mathcal{Q}}_\rho$. Since ρ is not estimable under H_0 (Davies, 1987), we propose to construct an overall test based on the commonly used minimum p-value approach that enables us to combine information from $\{\hat{\mathcal{Q}}_\rho, \rho \in \Omega\}$, where Ω is a pre-specified range for ρ . Specifically, we use the test statistic $\hat{\mathcal{P}}_{\min} = \min_{\rho \in \Omega} \hat{\mathcal{P}}_\rho$ and calculate its corresponding p-value as $\mathcal{B}^{-1} \sum_{b=1}^{\mathcal{B}} I(\hat{\mathcal{P}}_{\min}^{*b} \leq \hat{\mathcal{P}}_{\min})$, where $\hat{\mathcal{P}}_\rho = \mathcal{S}_\rho(\hat{\mathcal{Q}}_\rho)$, $\hat{\mathcal{P}}_\rho^{*b} = \mathcal{S}_\rho(\hat{\mathcal{Q}}_\rho^{*b})$, $\mathcal{S}_\rho(x) = \mathcal{B}^{-1} \sum_{b=1}^{\mathcal{B}} I(\hat{\mathcal{Q}}_\rho^{*b} \geq x)$, and $\hat{\mathcal{P}}_{\min}^{*b} = \min_{\rho \in \Omega} \hat{\mathcal{P}}_\rho^{*b}$.

In practice, for a given dataset, the optimal choice of kernel is also unclear. When the underlying effects are linear, one would expect the linear kernel to work well while overly complex kernels may result in low power due to over-fitting. On the other hand, non-linear kernels such as the knowledge of the underlying functional form of $h_0(\mathbf{Z})$, the optimal kernel that maximizes the power is unknown. Here, we propose an omnibus testing procedure to synthesize information from multiple kernels; the resampling methods enable convenient estimation of the null distribution of the omnibus test statistic. Suppose there are M

candidate kernels and for $m = 1, \dots, M$, let $\hat{\mathcal{P}}_{min}^{(m)}$ denote the observed p-value and

$\left\{ \hat{\mathcal{P}}_{min}^{(m)*b} \right\}_{b=1, \dots, \mathcal{B}}$ be $\mathcal{B}$ realizations from the estimated null distribution of $\hat{\mathcal{P}}_{min}^{(m)}$ as described above. Then an omnibus test can be formed based on

$\hat{\mathcal{P}}_{\mathcal{F}}^{omni} = \mathcal{F} \left(\hat{\mathcal{P}}_{min}^{(1)}, \dots, \hat{\mathcal{P}}_{min}^{(M)} \right)$, whose null distribution can be approximated using the

empirical distribution of $\left\{ \mathcal{F} \left(\hat{\mathcal{P}}_{min}^{(1)*b}, \dots, \hat{\mathcal{P}}_{min}^{(M)*b} \right) \right\}_{b=1, \dots, \mathcal{B}}$, where $\mathcal{F}(x_1, \dots, x_M)$ is some pre-specified function used to combine the p-values. In practice, one may choose $\mathcal{F}(P_1, \dots, P_M) = \min(P_1, \dots, P_M)$, which corresponds to the frequently used minimum p-value strategy. Alternative functions such as $\mathcal{F}(P_1, \dots, P_M) = \sum_{m=1}^M \{-\log(P_m)\}^d$ for some $d > 0$ can be used to more smoothly combine the p-values and a higher order d will reflect larger weights on the kernel with smaller p-values.

4. Simulations

We have conducted extensive simulation studies to examine the performance of our proposed procedures using linear (K_L), Gaussian (K_G), and quadratic (K_Q) kernels. We also carried out the omnibus testing procedure that combines p-values from different choices of

kernels based on $\mathcal{F}(P_1, \dots, P_M) = \sum_{m=1}^M \{-\log(P_m)\}^2$ ($Omnilog^2$). For comparison, we also carried out the standard Wald test under the Cox model with variance estimated using the robust variance (L_{χ^2}), the standard Wald test (L_{χ^2}) and the Wald test that includes both linear and quadratic effects (Q_{χ^2}). All results are summarized with 2000 simulated datasets for size estimation and 500 for power. The resampling was performed 1000 times for each dataset to estimate the null distributions.

We generated $\mathbf{U} = (U_1, U_2)^T$ with $U_1 \sim N(0, 1)$ and $U_2 \sim \text{Uniform}(0, 1)$. The new markers $\mathbf{Z}_{p \times 1}$ were generated from multivariate normal with zero-mean and a compound symmetric covariance matrix $\Sigma = (\mathbf{1} - \mathbf{r}_0) \mathbb{I}_{p \times p} + \mathbf{r}_0$, where $\mathbb{I}$ is the identity matrix and r_0 is the correlation among $\mathbf{Z}$ chosen to be either 0.2 or 0.5. The survival time T was generated from

$$\log(T) = -h_0(\mathbf{Z}) - U_1 - U_2 + \epsilon, \quad \text{with } \epsilon \sim \text{extreme value distribution.}$$

Four choices of $h_0(\mathbf{Z})$ was considered to reflect different effects of $\mathbf{Z}$ on T : (i) null effect

with $h_0(\mathbf{Z}) = 0$; (ii) linear effects with $h_0(\mathbf{Z}) = 0.05 \sqrt{\frac{p}{5}} \sum_{j=1}^5 Z_j$; (iii) interactive effects with

$h_0(\mathbf{Z}) = 0.1 \sqrt{\frac{p}{5}} (Z_1^2 + Z_2^2 + Z_3 Z_4 + Z_4^2 + Z_5)$; and (iv) complex non-linear effects with

$h_0(\mathbf{Z}) = \sqrt{\frac{p}{5}} \left[\Phi(Z_1^2 + 0.5 Z_1) + \Phi(Z_1 Z_2 + 0.5 Z_2) + 1.25 \{ \sin(3 Z_3) + \sin(3 Z_4) + \sin(3 Z_5) \} \right]$; where $\Phi(\cdot)$ is the cumulative distribution function of the standard normal. We let p be 5 or 20 but note that $h_0(\mathbf{Z})$ only depends on the first 5 elements of $\mathbf{Z}$. The censoring was generated from exponential with mean 0.2, leading to about 70% of censoring for settings (i) and (ii) and about 60% of censoring for settings (iii) and (iv). The full cohort size N was set

as 2000 and 5000 but only results from $N = 2000$ are shown since similar patterns were observed for $N = 5000$. To select the CCH subsample, we created a stratification variable $\mathcal{J}$ taking values 0, 1, 2 and 3 if $\Phi(Z_1 + Z_2 - 1) \in [0, 0.25]$, $(0.25, 0.5]$, $(0.5, 0.75]$ and $(0.75, 1]$, respectively. Within each of the strata indexed by the values of $\mathcal{J}$ and δ , we randomly sample $50m^{I(\delta=0)}$ subjects to the CCH sub cohort for $m = 1$ or 3.

The empirical size and power at a type I error rate of 0.05 from various testing procedures are summarized in Figure 1. As shown in Figure 1(a), the sizes for the proposed tests are all maintained near the nominal level across different choices of kernels and settings. On the other hand, the tests based on all standard methods tend to be anti-conservative with substantially inflated type I error especially for $m = 1$ or $p = 20$. For L_R , the size is almost always inflated across suggesting employing the robust variance for inference, while reasonable for the univariate case by being conservative, could lead to misleading results for the multivariate setting. For L_{χ^2} , the size is inflated when $m = 1$ or $p = 20$. This is likely due to the inaccurate approximation of the χ_p^2 to the null distribution in small sample sizes. The size is worse for Q_{χ^2} which is a $2p$ degree of freedom test and the χ_{2p}^2 does not appear to approximate the null distribution well except for the setting with $m = 3$, $p = 5$ and $r_0 = 0.5$. These results suggest that the standard Wald tests do not perform well in maintaining correct size, especially when quadratic bases are included. Since L_R has significantly inflated type I errors across various settings, we excluded its power results from the remaining plots for conciseness.

Under setting (ii) with linear effects, the Cox_{KM} test with K_L has the highest power as expected and the use of complex kernels resulted in substantial power loss. On the other hand, since K_Q captures quadratic and interactive effects, the Cox_{KM} test with K_Q performs the best in setting (iii). Note that although the Q_{χ^2} test has very high power in setting (iii), the results cannot be trusted due to the fact that it fails to maintain the size under the null.

Although the functional space spanned by K_G , $\mathcal{H}_{K_G}$ is dense and should be able to provide adequate approximation to second order polynomial functions for some ρ , the resulting complexity of $\mathcal{H}_{K_G}$ might be high and consequently the Cox_{KM} test using K_G is less powerful compared to that with K_Q due to the high degree of freedom. For setting (iv), the test with K_G outperforms other kernels as the complex non-linear effects cannot be adequately approximated using K_Q . For both settings (iii) and (iv), the linear kernel performed poorly compared to non-linear kernels. For all settings, the omnibus tests performed quite well with power comparable to the Cox_{KM} test constructed with the corresponding optimal kernel under each of the settings. This suggests that the omnibus testing approach is effective in adaptively synthesizing information from different kernels.

5. Application

The proposed testing procedures are applied to the Danish Diet, Cancer and Health (DCH) study (Tjønneland et al., 2007; Jensen et al., 2008) to assess the importance of apolipoprotein (apo) C-III (apoCIII) in modulating the effects of lipoproteins on the risk of coronary heart disease (CHD). ApoCIII, a small pro inflammatory protein, resides on the surface of lipoprotein particles and strongly affects their metabolism. It has been previously

suggested that presence or absence of apoCIII differentiates apoB and HDL into subtypes with differential associations with the risk of CHD (Jensen et al., 2012). Here, we are particularly interested in the presence of apoCIII on four lipoprotein particles, total cholesterol in plasma (Chol), apoB, HDL, and apoE-containing HDL (HDL_E). Each of these 4 lipoprotein particles was subfractionated into those with and without apoCIII. We denote the fractions of Chol, apoB, HDL and HDL_E without apoCIII respectively by %Chol₋, %apoB₋, %HDL₋, %HDL_{E-}. Our goal is to test whether each of these 4 apoCIII fraction markers is associated with CHD risk after adjusting for known risk factors $\mathbf{U}$ consisting of age, gender, alcohol consumption, smoking status, level of education, systolic and diastolic blood pressure, diabetes status, and hypertension status.

The original DCH study cohort consists of a total of 57,053 subjects between ages of 50 and 64 whose biospecimens were obtained at baseline and stored for future research. Among them, a total of 55,440 subjects (26,112 men and 29,328 women) without prior diagnosis of cancer or CHD at baseline are included in our analysis. A CCH subcohort, consisting of 3,604 participants, was selected to have their lipoprotein related markers measured (Jensen et al., 2012). The CCH sampling, performed by stratifying on $\mathcal{J}$ =gender and δ , selects all CHD cases up to December 31, 2003 and a random subset of additional 885 male and 800 female participants without CHD. The number of selected and total subjects in different strata are shown in Table 1(a).

To test the association between each of the apoCIII markers and CHD risk adjusting for $\mathbf{U}$, we used the full cohort data to estimate the null model parameters and constructed Cox_{KM} tests with three different kernels: K_L , K_G and K_Q . An omnibus test combining information from testing with all three kernels is also performed. The p-values from these tests are shown in Table 1(b).

Except for %apoB₋, all novel markers appear to be associated with CHD risk. Interestingly, the association for %HDL₋ is only apparent when the K_G is used with a p-value of 0.026 compared to p-value of 0.315 with K_L , suggesting that the underlying effect of %HDL₋ on CHD risk may be highly non-linear. This result suggests the potential advantage of allowing for complex effects using non-linear kernels. When testing the overall effects of all four apoCIII markers, we obtained highly significant p-values based on our proposed test across all choices of kernels. When combining information from all kernels using the proposed omnibus test, we observe a significant association with a p-value of < 0.001 indicating the potential utility of these apoCIII markers for improving risk prediction of CHD.

To further investigate the non-linear effect of %HDL₋, we also obtained the KM estimate of $h_0(\cdot)$ as described in Section 2 with K_G , where $\varphi=0.99$ and ρ for K_G and the penalty parameter $\mathcal{J}$ in (6) were both selected via the cross-validation. The estimator $\hat{h}(\cdot)$, as shown in Figure 2, demonstrates that the effect function is highly non-linear, suggesting that while higher %HDL₋ is generally associated with lower risk of CHD, the risk reduction is not linear with moderate reduction for %HDL₋ in the lower percentiles and sharper decline in risk when the %HDL₋ increases from 50th to 80th percentile. The protective effect of %HDL₋ appears to diminish when the %HDL₋ exceeds its 95th percentile although the estimator may be unstable towards the tail.

6. Discussion

Considerable research has focused on developing valid inference procedures for the regression parameters under standard survival regression models with CCH data. However, making valid inference for more complicated functionals beyond regression parameters remains a major obstacle to adopting the design in practice. Researchers often caution against using traditional bootstrap procedures with the case-cohort design, since it tends to overestimate the variance of parameters of interest (Samuelsen et al., 2007).

In this paper, with a goal of constructing hypothesis testing procedures under a Cox kernel machine model, we proposed a simple perturbation resampling approach to approximate the distribution of IPW test statistics obtained from CCH studies. Such a resampling approach is particularly useful for the present setting where explicit estimation of the null distribution is difficult. The key motivation behind our approach was based on the connection between the standard IPW estimator with true selection probabilities under $\mathbb{F}$ - sampling and the IPW estimators with estimated selection probabilities under $\mathbb{B}$ - sampling. Since the perturbed weights $\hat{\omega}_j^*$ are easy to calculate, the proposed procedures are easy to implement by repeatedly calculating $\hat{\omega}_j^*$ to obtain perturbed estimators. By letting $\mathcal{B}_j \sim \exp(1)$, the generation of $\hat{\omega}_j^*$ can be simplified by noting that

$$\pi_{dj}^* = \left\{ \sum_{j=1}^N V_j \mathcal{I}_{dj,i} \mathcal{B}_j \right\} / \left\{ \sum_{j=1}^N V_j \mathcal{I}_{dj,i} \mathcal{B}_j + \sum_{j=1}^N (1 - V_j) \mathcal{I}_{dj,i} \mathcal{B}_j \right\} \text{ and}$$

$$\sum_{j=1}^N (1 - V_j) \mathcal{I}_{dj,i} \mathcal{B}_j \sim \text{Gamma}(N_{dj} - n_{dj}, 1). \text{ Thus, one only needs to generate } \mathcal{B}_j \text{ for}$$

subjects in the CCH sub cohort as well as $2J$ independent Gamma random variables to calculate each set of $\hat{\omega}_j^*$.

Similar to the Cox_{KM} test for the full cohort data, the proposed Cox_{KM} test has the advantage of capturing non-linear effects and leveraging the potential correlation among the markers. One potential limitation of the proposed method is that it is computationally more demanding than the standard score test which can be performed using explicit variance estimators even under case cohort design. The range of the tuning parameter ρ for the kernel function plays an important role in the performance of the resulting test. Across all numerical studies, we chose ρ to achieve certain eigenvalue decay rate based on the estimated eigenvalues similar to those considered in Sinnott and Cai (2013). More rigorous rules for choosing the range of ρ warrant future research. When interest lies in estimation, adopting the kernel PCA approach can substantially improve both the estimation and computational efficiency. Further investigation is needed to identify an optimal approach to the simultaneous selection of the tuning parameters including both ρ , the ridge penalty ϑ , as well as the proportion of variance explained φ .

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

Payne and Neykov contributed equally to this work. This research was partially supported by NIH grants R01 GM079330, R01 GM085047, U54 H6007963, and R01 HL089778. The Diet, Cancer and Health Study is supported by the Danish Research Council and the Danish Cancer Society. HDL apoC-III measurements were funded by An Elite Research Award from the Danish Research Council for Independent Research.

References

- Anderson GL, Manson J, Wallace R, Lund B, Hall D, Davis S, Shumaker S, Wang C-Y, Stein E, Prentice RL. Implementation of the women's health initiative study design. *Annals of epidemiology*. 2003; 13:S5–S17. [PubMed: 14575938]
- 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]
- Braun, M. PhD thesis. University of Bonn; 2005. Spectral Properties of the Kernel Matrix and their Application to Kernel Methods in Machine Learning..
- Breslow NE, Wellner JA. Weighted likelihood for semiparametric models and two-phase stratified samples, with application to cox regression. *Scand. J. Statist*. 2007; 34:86–102.
- Cai T, Tonini G, Lin X. Kernel machine approach to testing the significance of multiple genetic markers for risk prediction. *Biometrics*. 2011; 67:975–986. [PubMed: 21281275]
- Colditz G, Manson J, Hankinson S. The Nurses' Health Study: 20-year contribution to the understanding of health among women. *Journal of Women's Health*. 1997; 6:49–62.
- Cristianini, N.; Shawe-Taylor, J. An Introduction to support vector machines: and other kernel-based learning methods. Cambridge University Press; Cambridge, UK: 2000.
- Davies R. Hypothesis testing when a nuisance parameter is present only under the alternative. *Biometrika*. 1987; 74:33–43.
- Gray RJ. Weighted analyses for cohort sampling designs. *Lifetime Data Analysis*. 2009; 15:24–40. [PubMed: 18712477]
- Gunter MJ, Hoover DR, Yu H, Wassertheil-Smoller S, Manson JE, Li J, Harris TG, Rohan TE, Xue X, Ho GY, et al. A prospective evaluation of insulin and insulin-like growth factor-i as risk factors for endometrial cancer. *Cancer Epidemiology Biomarkers & Prevention*. 2008; 17:921–929.
- Huang Y. Bootstrap for the case-cohort design. *Biometrika*. 2014; 101:465–476. [PubMed: 25170175]
- Jensen MK, Chiuve SE, Rimm EB, Dethlefsen C, Tjønneland A, Joensen AM, Overvad K. Obesity, behavioral lifestyle factors, and risk of acute coronary events. *Circulation*. 2008; 117:3062–3069. [PubMed: 18541738]
- Jensen MK, Rimm EB, Furtado JD, Sacks FM. Apolipoprotein C-III as a potential modulator of the association between hdl-cholesterol and incident coronary heart disease. *Journal of the American Heart Association*. 2012; 1:jah3–e000232. [PubMed: 23130121]
- Kimeldorf GS, Wahba G. A correspondence between bayesian estimation on stochastic processes and smoothing by splines. *Ann. Math. Statist*. 1970; 41:495–502.
- Kong L, Cai J, Sen R. Weighted estimating equations for semiparametric transformation models with censored data from a case-cohort design. *Biometrika*. 2004; 91:305–19.
- Li H, Luan Y. Kernel Cox regression models for linking gene expression profiles to censored survival data. *Pacific Symposium on Biocomputing*. 2003:65–76. [PubMed: 12603018]
- Liu D, Cai T, Zheng Y. Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics*. 2012; 68:1219–1227. [PubMed: 23173848]
- Nan B, Kalbfleisch J, Yu M. Asymptotic theory for the semiparametric accelerated failure time model with missing data. *Annals of Statistics*. 2009; 37:235–237.
- Prentice R. A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika*. 1986; 73:1–11.

- Rasmussen, CE.; Williams, CK. Gaussian processes for machine learning. MIT Press; Cambridge, Massachusetts: 2006.
- Samuelson SO, Ånestad H, Skrandal A. Stratified case-cohort analysis of general cohort sampling designs. *Scandinavian Journal of Statistics*. 2007; 34:103–119.
- Schölkopf, B.; Smola, AJ. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond (Adaptive Computation and Machine Learning)*. MIT Press; Cambridge, Massachusetts: 2001.
- Sinnott JA, Cai T. Omnibus risk assessment via accelerated failure time kernel machine modeling. *Biometrics*. 2013; 69:861–873. [PubMed: 24328713]
- Tjønneland A, Olsen A, Boll K, Stripp C, Christensen J, Engholm G, Overvad K. Study design, exposure variables, and socioeconomic determinants of participation in diet, cancer and health: a population-based prospective cohort study of 57,053 men and women in denmark. *Scandinavian Journal of Public Health*. 2007; 35:432–441. [PubMed: 17786808]
- Wacholder S, Gail M, Pee D, Brookmeyer R. Alternative variance and efficiency calculations for the case-cohort design. *Biometrika*. 1989; 76:117–123.

Figure 1.

Size and power of the proposed Cox_{KM} tests with K_L , K_G , and K_Q kernels; as well as the omnibus tests with $\mathcal{F}(P_1, \dots, P_3) = \sum_{m=1}^3 \{-\log(p_m)\}^2 (Omnilog^2)$. Shown also are the results based on standard Wald test procedures with linear effects and robust variance estimator (L_R , size only), linear effects (L_{χ^2}) and quadratic effects (Q_{χ^2}).

Figure 2.
Estimated log-hazard ratio function for %HDL_ under a Cox_{KM} model.

Table 1

Assessing the associations between %Chol₋, %apoB₋, %HDL₋, %HDL_{E-} and CHD risk with the DCH study.

(a) Sampling Design				
	Female, $\delta=0$	Female, $\delta=1$	Male, $\delta=0$	Male, $\delta=1$
Number of Available Subjects	28806	522	24715	1397
Number of Selected Subjects	800	522	885	1397
Sampling Fraction	0.028	1.0	0.036	1.00

(b) p-values based on the proposed KM tests with linear (K_L), Gaussian (K_G) and quadratic (K_Q) kernels. Shown also are the omnibus test results with $\mathcal{F}(P_1, P_2, P_3) = \sum_{m=1}^3 \{\log(P_m)\}^2$.				
	K_L	K_G	K_Q	Ominibus
%Chol ₋	0.028	0.035	0.248	0.047
%HDL _{E-}	0.000	0.003	0.002	0.000
%apoB ₋	0.860	0.607	0.822	1.000
%HDL ₋	0.315	0.026	0.661	0.130
ALL	0.000	0.002	0.002	0.000