Published in final edited form as:
*J Am Stat Assoc*. 2013 January 1; 108(504): 1532–1544. doi:10.1080/01621459.2013.856715.

# Resampling Procedures for Making Inference under Nested Case-control Studies

**Tianxi Cai**
Department of Biostatistics, Harvarfad University, Boston, MA, USA

**Yingye Zheng**
Public Health Sciences Division, Fred Hutchinson Cancer Research Center, Seattle, WA, USA

## Abstract

The nested case-control (NCC) design have been widely adopted as a cost-effective solution in many large cohort studies for risk assessment with expensive markers, such as the emerging biologic and genetic markers. To analyze data from NCC studies, conditional logistic regression (Goldstein and Langholz, 1992; Borgan et al., 1995) and maximum likelihood (Scheike and Juul, 2004; Zeng et al., 2006) based methods have been proposed. However, most of these methods either cannot be easily extended beyond the Cox model (Cox, 1972) or require additional modeling assumptions. More generally applicable approaches based on inverse probability weighting (IPW) have been proposed as useful alternatives (Samuelsen, 1997; Chen, 2001; Samuelsen et al., 2007). However, due to the complex correlation structure induced by repeated finite risk set sampling, interval estimation for such IPW estimators remain challenging especially when the estimation involves non-smooth objective functions or when making simultaneous inferences about functions. Standard resampling procedures such as the bootstrap cannot accommodate the correlation and thus are not directly applicable. In this paper, we propose a resampling procedure that can provide valid estimates for the distribution of a broad class of IPW estimators. Simulation results suggest that the proposed procedures perform well in settings when analytical variance estimator is infeasible to derive or gives less optimal performance. The new procedures are illustrated with data from the Framingham Offspring Study to characterize individual level cardiovascular risks over time based on the Framingham risk score, C-reactive protein (CRP) and a genetic risk score.

### Keywords

*Biomarker study*; *Interval Estimation*; *Inverse Probability Weighting*; *Nested case-control study*; *Resampling methods, Risk Prediction*; *Simultaneous Confidence Band*; *Survival Model*

## 1. INTRODUCTION

Accurate and individualized risk prediction is a crucial step in the path toward personalized medicine. To understand disease risks in a population, many large prospective cohorts have been assembled over the past decade in which patients were followed over time to observe clinical conditions. Unlike traditional case-control studies, such prospective cohort studies can provide valuable information on the absolute disease risk and the clinical utility of new novel markers for risk prediction. To enable future molecular biomarker studies, biological specimens of the study participants are often collected at baseline and stored. Once new markers emerge as promising predictive markers for a particular disease, stored specimens can be retrieved to obtain measurements of these markers. The stored samples collected from the full cohort are precious and must be used efficiently. Additionally due to costs associated with marker measurements, often it is undesirable and/or infeasible to obtain marker values for the entire study population. Two-phase study designs, including the case cohort and nested case–control (NCC) designs (Prentice and Breslow, 1978; Prentice, 1986; Thomas, 1977), have been proposed as cost-effective alternatives to the standard full-cohort design. In particular, under the NCC design, expensive markers are only measured for cases (those who developed events during follow up) and a set of controls randomly sampled from the risk sets of the cases *without replacement*. Controls are sometimes also matched to cases on key variables such as age and the batches of biological samples. The NCC design has been adopted to study biomarkers in many well known large cohort studies such as the Nurse's Health Study, Health Professional Follow-up Study, Physician's Health Study and Women's Health Initiative, and on a wide variety of complex diseases such as rheumatoid arthritis, type II diabetes, breast cancer, colorectal cancer, prostate cancer and coronary heart disease(Karlson et al., 2009; Hu et al., 2004; Ishibe et al., 1998; Hankinson et al., 1998; Ma et al., 1997; Wu et al., 2007; Nomura et al., 2000; Pradhan et al., 2002; Chlebowski et al., 2007).

The outcome dependent missingness in the NCC design generates complex data structure and thus imposes additional challenges for analysis. Statistical methods for analyzing data from NCC studies have developed for only limited settings. Under a Cox model (Cox, 1972), the most commonly used approach for estimating hazard ratio parameters is to fit a conditional logistic regression model with each stratum consisting of a case and individually matched controls (Goldstein and Langholz, 1992). For making inference about the absolute risk, a simple reweighted estimator focusing on within stratum comparisons has been developed (Borgan and Langholz, 1993; Borgan et al., 1995; Langholz and Borgan, 1997), with regression parameters estimated via the conditional logistic regression. In addition to its simplicity, conditional logistic regression based estimators also have the advantage of controlling for batch effects of matched case/control sets (Rundle et al., 2005). However, the conditional logistic regression estimators are only applicable to the Cox model and could be inefficient in certain settings with large relative risk parameter values (Scheike and Juul, 2004; Cai and Zheng, 2012). When the batch effects are negligible, one may break the matching and consider alternative estimators. For example, Scheike and Juul (2004) proposed a maximum likelihood estimator (MLE) which attains the semi-parametric efficiency bound. Saarela et al. (2008) extended the MLE to the competing risks setting. Zeng et al. (2006) extended the MLE to the more general class of semi-parametric transformation models. One major limitation of the MLE is that it requires the estimation of the conditional density of the new markers given other clinical variables, which may be infeasible when the dimensions of new markers and clinical variables are not small unless additional modeling assumptions are imposed. In addition, the MLE is only valid when the censoring distribution is independent of the new markers. Recently, inverse probability weighted (IPW) estimators have been developed as useful alternatives to make inference for various models and parameters (Samuelsen, 1997; Chen, 2001; Samuelsen et al., 2007; Cai and Zheng, 2011, 2012; Lu and Liu, 2012). Furthermore, Unlike the conditional logistic regression based estimators and the MLE, the IPW approach (i) can be easily extended beyond the Cox model to other modeling paradigms such as the accelerated failure time model, (ii) can facilitate the calculation of parameters beyond relative and absolute risks (Cai and Zheng, 2012), (iii) can accommodate settings with multiple selectively measured markers, and (iv) allow for easy combination of information from previous studies and hence leads to more efficient design with respect to cost in measurement (Salim et al., 2009).

An IPW estimator $\widehat{\theta}$ is often obtained as the minimizer of a weighted objective function $n^{-1}\sum_{i=1}^{n}\widehat{w_i}L\left(\theta;D_i\right)$, where $L$ is a pre-specified loss function, $D^i$ is the vector of observations from the *i*th subject, $\widehat{w_i}=V_i/\widehat{p_i}$ is the IPW weight for the *i*th subject with

$E(\widehat{w_i}|D_i)=1$, $V_i$indicates whether the *i*th subject is sampled into the NCC sub-cohort and $\widehat{p_i}$ is the probability of this subject being selected into the NCC subcohort. Conditional on the observed event times, the risk sets are of finite size. Sampling *without replacement* from such finite populations induces a complex correlation structure among $V_i$'s and thus brings difficulty to inference. Naive variance estimates without accounting for the correlation may lead to substantial over-estimation. In cases where $L$is smooth in $\theta$, analytic forms of variance estimators that account for such a correlation could be derived and used for pointwise interval estimation. However there are several reasons alternative inference procedures are needed. In practice it may be numerically burdening to derive analytic variance estimators on a case by case basis, particularly when parameters of interest are complex functions involving not only $\hat{\theta}$ but some distributional functions. Additionally, in biomarker evaluation, to make decision regarding optimal time points or biomarker cutoff values for future predictions, one needs to draw simultaneous inference over a range of time points or marker values. Such simultaneous inference typically involves the estimation of the joint distribution of a process, which is often not tractable explicitly. Furthermore, in many situations such as when $L$is not smooth in $\theta$or involves regularization in the presence of high dimensional covariates, explicit variance estimation might not be feasible and/or reliable. With standard cohort studies, one may rely on resampling procedures such as the bootstrap to make inference for such settings. Unfortunately, standard bootstrap procedures cannot be directly applied here due to the correlation among $V_i$'s. Modified bootstrap procedures proposed to account for finite population sampling from survey sampling literature (Rao and Wu, 1988; Rao et al., 1992; Sitter, 1992; Shao, 2003) only accommodate a relatively simple correlation structure. Existing modified bootstrap methods for analyzing time series or spatially correlated data often require stationary assumptions and rely heavily on the specific correlation structure under those settings (Garcia-Soidan and Hall, 1997; Hall et al., 1995; Härdle et al., 2003). Unfortunately, these existing approaches cannot be applied to the NCC setting where the sampling is performed repeatedly over the risk sets defined by the event times of all the cases, which induced a much more complex dependent structure in data. Developing a general inference tool for risk modeling involving biomarkers sampled under the NCC design would be of great practical value.

In this paper, we propose perturbation resampling methods for approximating the distribution of IPW estimators with NCC studies. These tools are particularly useful for the aforementioned settings where simultaneous inference over a curve is of interest or analytical derivation is impractical or infeasible. We provide justification for our resampling procedure by making a connection between the IPW estimators constructed with (i) true selection probabilities under the standard finite population (F) sampling when controls are sampled *without replacement*; and (ii) estimated selection probabilities under the Bernoulli (B) sampling when controls are sampled *with replacement* from the risk sets. Although the perturbation cannot directly mimic the correlation among the $V_i$'s under F-sampling, we recover the effect of the correlation on the variance by estimating selection probabilities for each perturbed sample. We detail the resampling procedures in Section 2 and illustrate these procedures using the Cox model and an accelerated failure time (AFT) model in Section 3. Results from simulation studies suggest that the resampling procedures work well in finite sample. We report simulation results and an application of our procedures to the Framingham Offspring Study in section 4. Some concluding remarks are given in Section 5. Theoretical justifications for the resampling method are outlined in the Appendix.

## 2. RESAMPLING BASED INFERENCE PROCEDURES FOR IPW ESTIMATORS

### 2.1 IPW Estimators for NCC Data

Suppose we have a cohort of $n$ individuals followed prospectively for an event. Because of censoring, the event time $T$ is only observable up to a bivariate vector ($X$, $\delta$), where $X = T \wedge C$, $\delta = I(T \leq C)$ and $C$ is the censoring time assumed to have a finite support [0, τ]. Let $\mathscr{D}$= {$\mathrm{D_i} = (X_\mathrm{i},\delta_\mathrm{i},\mathrm{W_i}')'$, $i$ = 1,...,$n$} denote the full cohort data, where $\mathbf{W}_i$ denotes markers only measured if the $i$th subject was selected into the NCC subcohort, denoted by $\mathscr{D}_{\mathrm{NCC}}$. We consider the prediction of survival up to $\tau_0 < \tau$ such that $\inf_{\mathbf{W}} P(X > \tau_0 \mid \mathbf{W}) > 0$.

Without loss of generality, we assume that all cases are included in $\mathscr{D}_{\mathrm{NCC}}$. Controls are selected randomly from the risk set of cases. For example, for a case with failure time $X_i$, $m$ controls are randomly sampled *without replacement* from the risk set of $X_i$ excluding the case, denoted by $\mathscr{R} = \{k : 1 \leq k \leq n, X_k \geq X_i, k \neq i\}$. Let $V_{0ij}$ be a binary random variable with $V_{0ij} = 1$ denoting the $j$th subject being selected as a control for the $i$th subject. Then $\mathscr{D}_{\mathrm{NCC}}$ consists of all subjects with $V_j = \delta_j + (1 - \delta_j)V_{0j}$ taking value 1, where $V_{0j}{=}1-\Pi_{i:j\varepsilon\mathscr{R}_i}$ indicates the $j$th subject being sampled into $\mathscr{D}_{\mathrm{NCC}}$ as a control. The corresponding sampling probability is $\widehat{p_j}{=}P\left(V_j{=}1|\mathscr{D}\right){=}\delta_j{+}\left(1-\delta_j\right)\widehat{p_{0j}}$, where $\widehat{p}_j{=}P\left(V_j{=}1|\mathscr{D}\right){=}\delta_j{+}\left(1-\delta_j\right)\widehat{p}_{0j}$, and $\widehat{p}_{0j}{=}P\left(V_{0j}{=}1|\mathscr{D}\right){=}1{-}\widehat{G}\left(X_j\right),\widehat{G}\left(t\right){=}\Pi_{i:X_i\leq t}\left\{1-m\delta_i/\|\mathscr{R}_i\|\right\}, \quad \text{and} \quad \|\mathscr{R}_i\|{=}\sum_{k=1}^{n}I\left(X_k\geq X_i-1\right)$ (Samuelsen, 1997). To account for missingness of $Y$ under 𝔽sampling, one may follow the general IPW principle and weight the $j$th subject by $\widehat{w}_j{=}V_j/\widehat{p}_j{=}\delta_j{+}\left(1-\delta_j\right)V_{0j}/\widehat{p_{0j}}$.

### 2·2 Resampling Methods to Approximate the Distribution of IPW Estimators

To make inference about IPW estimators with NCC data, it is important to recognize that the $V_i$'s are weakly correlated. From Samuelsen (1997), we have

$$\widehat{r}_{ij}{=}\mathrm{cov}\left(\widehat{w}_i,\widehat{w}_j|\mathscr{D}\right){=}-n^{-1}m\int\eta\left(t;X_i,\delta_i\right)\eta\left(t;X_j,\delta_j\right)\frac{d\Lambda_{\mathrm{marg}}(t)}{\pi(t)}{+}O_p\left(n^{-3/2}\right),$$

$$\begin{aligned}
&\eta\left(t;X_j,\delta_j\right)\\
&\quad{=}I\left(X_j{>}t\right)\left(1-p_j\right)\\
&\quad/p_j,\Lambda_{\mathrm{marg}}\left(t\right)\\
&\quad{=}\int_0^t\pi(u)^{-1}dA_{\mathrm{marg}}\left(u\right)\pi\left(t\right)\\
&\quad{=}P\left(X_j\geq t\right),p_j\\
&\quad{=}\delta_j{+}\left(1-\delta_j\right)\left\{1-G\left(X\right)\right\},A_{\mathrm{marg}}\left(t\right)\\
&\quad{=}E\left\{N_i\left(t\right)\right\},N_i\left(t\right)\\
&\quad{=}\delta_iI\left(X_i\leq t\right),\quad\mathrm{and}G\left(t\right)
\end{aligned}$$

where $=\exp\left\{-m\Lambda\left(t\right)\right\}$. It is not difficult to see that the $n^{-1}$ rate of covariance is not negligible at the first order when calculating the asymptotic variance of the weighted estimators.

For illustration, consider a generic IPW estimator $\widehat{\theta}_{t_0}$, which is a minimizer of an IPW objective function $\widehat{L}_{t_0}\left(\theta\right){=}n^{-1}\sum_{j=1}^{n}\widehat{w}_jL_{t_0}\left(\theta;\mathrm{D}_\mathbf{j}\right)$ where $L_{t0}$ ($\theta$; **D**) is a given loss function that is continuous in $\theta$ whose gradient function $\mathbf{U}^{t0}$ ($\theta$; **D**) exists such that $E\left\{\mathrm{U}_{t_0}\left(\theta;\mathrm{D}_\mathbf{j}\right)\right\}{=}\partial E\left\{L_{t_0}\left(\theta;\mathrm{D}_\mathbf{j}\right)\right\}\partial\theta$, $\|\mathbf{E}\left\{\mathrm{U}_{\mathbf{t_0}}\left(\theta;\mathrm{D}_\mathbf{j}\right)\right\}\|_\mathbf{2}^\mathbf{2}{<}\infty$, and $\|\boldsymbol{a}\|_2$ denotes the $L_2$

norm of vector ***a***. Details of the regularity conditions for the loss functions are given in Appendix A. Here, we let $\widehat{\theta}_{t_0}$ be possibly indexed by $t_0$ to accommodate the estimation of time-dependent parameters. Under suitable regularity conditions,

$$\widehat{W}_{t_0}=n^{\frac{1}{2}}\left(\widehat{\theta}_{t_0}-\theta_{t_0}\right)\approx\mathbb{A}_{t_0}(\theta_{t_0})^{-1}\widehat{W}_{\mathrm{U}_{t_0}},\quad\widehat{W}_{\mathrm{U}_{t_0}}$$

$$=n^{-\frac{1}{2}}\sum\nolimits_{j=1}^{n}\widehat{w}_j\mathrm{U}_{t_0}\left(\mathrm{D}_j\right)$$

$$=\mathrm{U}_{t_0}\left(\theta_{t_0};\mathrm{D}_j\right)\quad\mathbb{A}_{t_0}\left(\theta\right)$$

$$=\delta^2E\left\{L_{t_0}\left(\theta;\mathrm{D}_{\mathbf{j}}\right)\right\}/\partial\theta\partial\theta^{'}$$

, $\mathbf{U}^{t0}$ $(\mathbf{D}_j)$ = $\mathbf{U}^{t0}$ $(\theta_{t0}$; $\mathbf{D}_j)$ and $\mathbb{A}_{t_0}\left(\theta\right)=\partial^2E\left\{L_{t_0}\left(\theta;\mathrm{D}_{\mathbf{j}}\right)\right\}/\partial\theta\partial\theta^{'}$. Hence, the asymptotic variance of $\widehat{W}_{t_0}=n^{\frac{1}{2}}\left(\widehat{\theta}_{t_0}-\theta_{t_0}\right)$ is $\sum_{t_0}=\mathbb{A}_{t_0}(\theta)^{-1}\mathbb{C}_{\mathrm{U}_{t_0}}\mathbb{A}_{t_0}(\theta_{t_0})^{-1}$, where

$$\mathbb{C}_{\mathrm{U}_{t_0}}=\check{\mathbb{C}}_{\mathrm{U}_{t_0}}-\int\eta_{\mathrm{U}_{t_0}}(t)^{\otimes 2}\pi(t)^{-1}d\Lambda_{\mathrm{marg}}\left(t\right),\quad\check{\mathbb{C}}_{\mathrm{U}_{t_0}}=E\left\{\mathrm{U}_{t_0}(\mathrm{D}_j)^{\otimes 2}\left(1-p_j\right)/p_j\right\},\quad(2.1)$$

$\eta_{\mathrm{U}_{t_0}}\left(t\right)=E\left\{I\left(X_j>t\right)\left(1-p_j\right)/p_j\mathrm{U}_{t_0}\left(D_j\right\}\right.$, for any vector ***a***, $\boldsymbol{a}^{\otimes 1}$ = ***aa′***. The term $\mathbb{C}\mathbf{U}t_0$ corresponds to the *robust variance* when the correlation among the $V_j$ ’s is ignored or when the controls were selected via $\mathbb{B}$sampling. After correcting for the correlation due to $\mathbb{F}$ sampling, $\mathbb{C}_{\mathbf{U}t0}$ is always no greater than $\check{\mathbb{C}}_{\mathrm{U}_{t_0}}$ and ignoring the correlation would often result in over-estimation of $\Sigma_{t0}$.

When the parameter of interest does not vary over $t_0$ and the objective function is smooth in $\theta$, it is not difficult to estimate $\Sigma_{t0}$ explicitly. For example, empirical variance estimator for the hazard ratio parameter under the Cox model can be constructed easily as in Samuelsen (1997) and Cai and Zheng (2012). However, such explicit estimation may be infeasible for more complex settings such as making simultaneous inference about a curve or estimations involving non-smooth objective functions. The construction of simultaneous confidence intervals (CIs) often requires approximating the distribution of a Gaussian process whose covariance function may be difficult to estimate explicitly. When the objective function is not smooth, explicit estimation of $\mathbb{A}^{t0}$ $(\theta_{t0})$ may not be feasible. For these settings, resampling methods have been proposed as a useful tool for inference (Davison et al., 1999; Jin et al., 2001; Park and Wei, 2003). However, it is well known that standard bootstrap methods fail to approximate the distribution of $\widehat{W}_{t_0}$ under two-phase design with $\mathbb{F}$sampling (Gray, 2009). For example, a naive bootstrap of the observations in $\mathscr{D}_{\mathrm{NCC}}$ with $\widehat{w}_j$ fixed destroys the correlation among the $V_j$ ’s due to $\mathbb{F}$sampling and leads to the robust variance $\check{\Sigma}_{t_0}=\mathbb{A}_{t_0}(\theta_{t_0})^{-1}\check{\mathbb{C}}_{\mathrm{U}_{t_0}}\mathbb{A}_{t_0}(\theta_{t_0})^{-1}$. To mimic the effect of the correlation, we propose a resampling procedure motivated by the equivalence between $\widehat{W}_{\mathrm{U}_{t_0}}=n^{-\frac{1}{2}}\sum_{j=1}^{n}\widehat{w}_j\mathrm{U}_{t_0}\left(\mathrm{D}_j\right)$ under $\mathbb{F}$sampling and $\widehat{W}_{\mathrm{U}_{t_0}}^{B}=n^{-\frac{1}{2}}\sum_{j=1}^{n}\widehat{w}_j^{B}\mathrm{U}_{t_0}\left(\mathrm{D}_j\right)$ under $\mathbb{B}$sampling, where

$$\widehat{w}_j=\delta_j+\left(1-\delta_j\right)V_{0j}/\widehat{\widehat{p}}\quad\text{with}\quad\widehat{\widehat{p}}_{0j}=1-\prod_{i:j\in\mathscr{R}_i}\left\{1-\frac{\sum\limits_{l\in\mathscr{R}_i}V_{0il}\delta_i}{||\mathscr{R}_i||}\right\}\quad(2.2)$$

Under $\mathbb{B}$sampling, $V_{0jl}$’s are independent of each other conditional on $\mathscr{D}$and $P\left(V_{0j}=1|\mathscr{D}\right)=\widehat{p_{0j}}$. In (2·2), the true selection probability $\widehat{p_{0j}}$ in $\widehat{w}_j$ is replaced with an estimate $\widehat{\widehat{p}}_{0j}$ using observed $V_{0jl}$’s. As shown in Cai and Zheng (2011) that the asymptotic variance of $\widehat{p}_{0j}$ under $\mathbb{B}$sampling is also $\mathbb{C}_{\mathbf{U}t0}$. Note that a similar result has been previously shown in Breslow and Wellner (2007) for IPW estimators under stratified case-cohort design. This equivalence suggests that the variance correction due to the correlation among $V_j$'s under $\mathbb{F}$sampling can be recovered by perturbing $V_j$ 's as if they were independent but inducing a correlation by using *estimated* selection probabilities obtained from the perturbed $V_j$'s. Therefore, the perturbation method can recover the variance of IPW estimators obtained under $\mathbb{F}$sampling. If $\mathbb{B}$sampling was employed for the study design and the IPW estimators are constructed based on estimated weights, then the proposed perturbation procedures can also be used to approximate the variance of such IPW estimators.

To approximate the distribution of $\widehat{W}_{t_0}$ via perturbation resampling, we propose the following resampling procedure:

1. Generate $n^2$ random realizations of $I_{il}$ from a known distribution with $E(I_{il})$ and create $I = \{I_{il}, i = 1,...,n; l = 1,...,n\}$.

2. Use $I$ to obtain perturbed weights $\widehat{w}_j^* = V_j^* / \widehat{p}_j^*$, where $V_j^* = \delta_j I_{jj} + (1-\delta_j) V_{0j}^*, \widehat{p}_j^* = \delta_j + (1-\delta_j) \widehat{p}_{0j}^*, \widehat{p}_{0j}^* = 1 - \exp\left\{-\widehat{\Lambda}_{\mathrm{marg}}^* (X_j)\right\}$.

$$V_{0j}^* = 1 - \prod_{i:j\in\mathscr{R}_i} (1 - \delta_i V_{0ij} I_{ij}), \quad \widehat{\Lambda}_{\mathrm{marg}}^*(t) = \sum_{i:X_i\le t, \delta_i=1} \frac{\sum_{l\in\mathscr{R}_i} V_{0il} I_{il}}{||\mathscr{R}_i||},$$

3. Use the perturbed weights to obtain the perturbed counterpart of $\widehat{L}_{t_0}(\theta)$ as

$$\widehat{L}_{t_0}^*(\theta) = n^{-1} \sum_{j=1}^{n} \widehat{w}_j^* L_{t_0}(\theta; \mathrm{D}_{\mathbf{j}})$$

   and let $\widehat{\theta}_{t_0}^* = \mathrm{argmin}_\theta \widehat{L}_{t_0}^*(\theta)$.

4. Repeat steps 1-3 for $B_0$ times to obtain $B_0$ realizations of $\widehat{\theta}_{t_0}^*$, denoted by $\left\{\widehat{\theta}_{t_0}^{(b)}, b=1, ..., B_0\right\}$. The empirical distribution of $\widehat{W}_{t_0}^{(b)} = n^{\frac{1}{2}}\left(\widehat{\theta}_{t_0}^{(b)} - \widehat{\theta}_{t_0}\right)$ conditional on the observed data can be used to approximate the distribution of $\widehat{W}_{t_0}$. For example, the variance of $\widehat{W}_{t_0}$ can be approximated by $B_0^{-1}\sum_{b=1}^{B0} \widehat{W}_{t0}^{(b)} \widehat{W}_{t_0}^{(b)\prime}$.

In Appendix A, we provide the justification for why the distribution of $\widehat{W}_{t_0}^* = n^{\frac{1}{2}}\left(\widehat{\theta}_{t_0}^* - \widehat{\theta}_{t_0}\right)$ given the observed data $\mathscr{F} = \{X_i, \delta_i, \mathbf{Z}_i, \mathrm{V}_i\mathbf{Y}_i, V_{0ij}I(X_i \le X_j)\delta_i, i, j = 1,..., n\}$ can be used to approximate the unconditional distribution of $\widehat{W}_{t_0}$. It is important to note that $\{I_{il}\}$ is only used in the construction of $\widehat{W}_{t_0}$ when $\delta_i = 1$ if $i = l$, or $\delta_i(1 - \delta_l)V_{0il} = 1$ if $i \le l$. Thus, to compute $\widehat{W}_{t_0}^*$, only $(m + 1)n_1 2$ of $\{\mathrm{I}_{il}\}$ need to be generated, with $n_1 = \Sigma_i\, \delta i$. For computational ease, one may replace $\widehat{w}_j^*$ in the above equations with $\widehat{w}_j^* \bigwedge 0$ since $V_{0j}{}^*$ may be negative when $\sum_{i=1}^{n} \delta_i V_{0ij} \ge 2$, i.e. when the $j$th subject is sampled as a control for multiple cases. However, since $P\left(\sum_{i=1}^{n} \delta_i V_{oij} \ge 2\right) = O_p\left(n^{-1}\right)$, the truncation of $\widehat{\ }_j^*$ is negligible at the first order in large sample since

$n^{\frac{1}{2}}\sum_{j=1}^{n}|\widehat{w}_j^*|I\left(\widehat{w}_j^*<0\right) \leq n^{\frac{1}{2}}\sum_{j=1}^{n}|\widehat{w}_j^*|I\left(\sum_{j=1}^{n}\delta_i V_{0ij} \geq 2\right) =_{O_p}(1)$. To adjust for the effect of truncation in finite sample, one may adjust the SE by a factor of $\widehat{c}=E\left\{\sum_{j=1}^{n}|\widehat{w}_j^* - \widehat{w}_j/\sum_{\widehat{w}_j^*>0}|\widehat{w}_j^* - \widehat{w}_j\right\}|\mathscr{F}\ \}$ and it is straightforward to see that $\widehat{c} \rightarrow 1$ in probability.

### 2·3 Resampling Procedures for IPW Estimators with Additional Matching

When controls are selected using additional matching variables **Z**, we may obtain $\widehat{p}_j$ and $\widehat{r}_{ij}$ accordingly. Without loss of generality, we assume that the eligible matching set for a case with **Z** = **z** consists of {$\mathbf{Z}_i$ : |$\mathbf{Z}_i$ − **z**| ≤ $\boldsymbol{a}_0$}, where $\boldsymbol{a}_0$ is a pre-specified constant that determines the range of matching and the inequality holds element-wise. The matched risk set for the *i*th subject is $\mathscr{R}_{\mathscr{I}}^*=\{k:1 \leq k \leq n, k \neq i, X_k \geq X_i, |\mathrm{Z}_k - \mathrm{Z}_i| \leq a_0, k \neq i\}$ and the IPW weights $\widehat{w}_j$ should be replaced by $\widehat{w}_j^{\dagger}=V_j^{\dagger}/\widehat{p}_j^{\dagger}$, where $V_j^{\dagger}=\delta_i+(1-\delta_i)V_{0j}^{\dagger}, V_{oj}^{\dagger}=1-\Pi_{i:j\epsilon\mathscr{R}_i^{\dagger}}(1-\delta_i V_{0ij}), \widehat{p}_j^{\dagger}=\delta_j+\left(1-\widehat{G}(X_j,\mathrm{Z}_j)\right)$ is the probability for the *j*th subject ever being sampled into $\mathscr{D}_{\mathrm{NCC}}$ under matching, $\widehat{G}(X_j,\mathrm{Z}_j)=\Pi_{i:j\epsilon\mathscr{R}_i^{\dagger}}\left\{1-m\delta_i/\|\mathscr{R}_i^{\dagger}\|\right\} \quad \text{and} \quad \|\mathscr{R}_i^{\dagger}\|=\sum_{k=1}^{n}I(X_k \geq X_i, |\mathrm{Z}_k - \mathrm{Z}_i \leq a_0)-1$. The asymptotic variance of $\widehat{W}_{t_0}$ is given in (A.4) of Appendix B. To approximate the distribution of $\widehat{W}_{t_0}$ in this setting, we replace $\widehat{w}_j^*$ in section 2·2 with $\widehat{w}_j^{\dagger}=V_j^{\dagger}/\widehat{p}_j^{\dagger}$, where $V_j^{\dagger *}=\delta_j I_{jj}+(1-\delta_j)V_{0j}^{\dagger *}, V_{0j}^{\dagger *}=1-\Pi_{i:j\epsilon\mathscr{R}_i^{\dagger}}(1-\delta_i V_{0ij}I_{ij}), \widehat{p}_j^{\dagger *}=\delta_j+(1-\delta_j)\left\{1-e^{-\widehat{\Lambda}_{\mathrm{marg}}^{\dagger *}(X_j,\mathrm{Z}_j)}\right\}$, and

$$\widehat{\Lambda}_{\mathrm{marg}}^{\dagger *}(X_j,\mathrm{Z}_j)=\sum_{i:j\in\mathscr{R}_i^{\dagger}}\delta_i\frac{\sum\limits_{l\in\mathscr{R}_i^{\dagger}}V_{0il}I_{il}}{||\mathscr{R}_i^{\dagger}||}.$$

.

## 3. ILLUSTRATIVE APPLICATIONS

### 3·1 Simultaneous Inference under the Standard Cox Model

Consider the Cox proportional hazards model Cox (1972),

$$S_{\mathrm{W}_0}(t_0)=P(T \geq t_0|W=W_0)=\exp\left\{\Lambda_0(t)e^{\beta'\mathrm{W}_0}\right\},$$

where $\Lambda_0(\cdot)$ is an un-specified baseline cumulative hazard function, and $\boldsymbol{\beta}$ is the unknown log hazard ratio parameter. We next illustrate how the resampling procedure can be used to make simultaneous inference about the survival function $\mathrm{S}^{\mathbf{W}0}$ ($t_0$) across $t_0$ ε $T$ = [$\tau_l$, $\tau_r$], where 0 < $\mathrm{S}^{\mathbf{W}0}$ ($\tau_r$) < $\mathrm{S}^{\mathbf{W}0}$ ($\tau_l$) < 1.

From Samuelsen (1997) and Cai and Zheng (2012), a consistent estimator for $S^{\mathbf{W0}}$ ($t_0$) may be obtained as

$$\begin{aligned}\widehat{S}_{\mathrm{W}_0}\left(t_{0)}\right) &= \exp\left\{-\widehat{\Lambda}_0(t)\exp\left(\widehat{\beta}'\mathrm{W}_0\right)\right\} \quad \text{where} \widehat{\beta} \\ &= \text{argmax}_{\beta}\widehat{l}(\beta), \widehat{\Lambda}_0(t) \\ &= \int_0^t \widehat{\Pi}_0\left(u,\widehat{\beta}\right)^{-1} d\widehat{N}(u), \widehat{N}(u) \\ &= \sum_{j=1}^n \widehat{w}_j N_j(u) \\ &/\sum_{j=1}^n \widehat{w}_j\end{aligned}$$

.

$$\widehat{l}(\beta) = \sum_{j=1}^{n} \widehat{w}_j \int \left\{\beta'\mathrm{W}_j - \log\widehat{\Pi}_0(t,\beta)\right\} dN_j(t), \quad \text{and} \quad \widehat{\Pi}_k(t,\beta) = \frac{\sum_{i=1}^{n} \widehat{w}_i \mathrm{W}_i^{\otimes k} I(X_i \geq t) e^{\beta' W_i}}{\sum_{i=1}^{n} \widehat{w}_i}$$

Furthermore, following Cai and Zheng (2012), for any given $\mathrm{W}_0$,

$\mathrm{W}_0, \widehat{W}_{t_0} = n^{\frac{1}{2}}\left\{\widehat{S}_{\mathrm{W}_0}(t) - S_{\mathrm{W}_0}\left(t_{0)}\right)\right\}$ converges weakly to a zero-mean Gaussian process in $t_0$, denoted by $W^{t0}$. For each fixed $t_0$, it is not difficult to estimate the distribution of $\widehat{W}_{t_0}$ explicitly based on large sample normal approximations. However, it is generally difficult to explicitly estimate the joint distribution of $\widehat{W}_{t_0}$ across $t_0$ ε $\mathscr{T}$ which is useful for constructing simultaneous CIs for {$S^{\mathbf{W0}}$ ($t_0$), $t_0$ ε $T$}. On the other hand, the proposed perturbation procedure provides a convenient solution to such a problem. Specifically, let

$$\widehat{\beta}^* = \text{argmax}_{\beta}\widehat{l}^*(\beta) \quad \text{and} \quad \widehat{\Lambda}_0^*(t) = \int \widehat{\Pi}_0^*(t,\beta^*)^{-1} d\widehat{N}^*(u), \quad \text{where} \quad \widehat{N}^*(u) = \sum_{j=1}^n \widehat{w}_j^* N_j(u) / \sum_{j=1}^n \widehat{w}_j^*$$

,

$$\widehat{l}^*(\beta) = \sum_{j=1}^{n} \widehat{w}_j^* \int \left\{\beta'\mathrm{W}_j - \log\widehat{\Pi}_0^*(t,\beta)\right\} dN_j(t) \quad \text{and} \quad \widehat{\Pi}_k^*(t,\beta) = \frac{\sum_{i=1}^{n} \widehat{w}_j^* I(X_j \geq t) \exp\left(\beta'\mathrm{W}_j\right) \mathrm{W}_j^{\otimes k}}{\sum_{i=1}^{n} \widehat{w}_j^*}.$$

.

Then a perturbed version of $\widehat{S}_{\mathrm{W}_0}(t_0)$ can be obtained as $\widehat{S}_{\mathrm{W}_0}^*(t_0) = \exp\left\{-\widehat{\Lambda}_0^*(t_0)\exp\left(\mathrm{W}_0'\widehat{\beta}^*\right)\right\}$. In Appendix C, we show that $n^{\frac{1}{2}}\left(\widehat{\beta}^* - \widehat{\beta}\right)$ given $\mathscr{F}$ can be used to approximate the distribution of $n^{\frac{1}{2}}\left(\widehat{\beta}^* - \widehat{\beta}\right)$ and for $t_0$ ε $\mathscr{T}$ $\widehat{W}_{t_0}^* = n^{\frac{1}{2}}\left\{\widehat{S}_{\mathrm{W}_0}^*(t_0) - \widehat{S}_{\mathrm{W}0}(t_0) - \widehat{S}_{\mathrm{W}_0}(t_0)\right\}$ given $\mathscr{F}$ converges weakly to $W^{t0}$. In practice, one may generate $B_0$ sets of $I$ to obtain the corresponding perturbed IPW weights $\left\{\widehat{w}_j^{(b)}, j=1, ..., n\right\}_{b=1,...B_0}$ and subsequently computing $B_0$ realizations of $\widehat{W}_{t_0}^*$ as $\left\{\widehat{W}_{t_0}^{(b)}\right\}_{b=1,...,B_0}$. Let $\widehat{\sigma}_{S\mathrm{W}_0}^2(t_0)$ ($t_0$) be the empirical variance of $\widehat{W}_{t_0}^{(b)}$, then a 100(1 − $\alpha$)% pointwise and simultaneous CIs for {$S^{\mathbf{W0}}$ ($t_0$), $t_0$ ε $\mathscr{T}$ may be respectively constructed as $\widehat{S}_{\mathrm{W}_0}(t_0) \pm Z_{1-\alpha/2}{}^{\widehat{\sigma}_{S\mathrm{W}_0}}(t_0)$ and $\widehat{S}_{\mathrm{W}_0}(t_0) \pm Q_{1-\alpha}{}^{\widehat{\sigma}_{S\mathrm{W}_0}}\, t_0$, where $\mathscr{Z}_{\gamma}$ and $\mathscr{Q}_{\gamma}$ are the 100$\gamma$th percentile of $N$ (0, 1) and $\left\{\sup_{t_0 \epsilon \mathscr{F}}\widehat{\sigma}_{S_{\mathrm{W}_0}}(t_0)^{-1}|\widehat{W}_{t_0}^{(b)}|, b=1, ..., B_0\right\}$, respectively. In practice, when $S^{\mathbf{W0}}$ ($t_0$) is close to 0 or 1, one may construct the CIs based on $g${$S^{\mathbf{W0}}$ ($t_0$)}, where $g(x)$ = log{− log($x$)}.

### 3·2 Inference under the AFT Model

The AFT model essentially assumes a linear regression model for log($T_i$):

$$\log\left(T_i\right)=\beta^{'}\mathrm{W}_i+\epsilon_i \quad \text{where} P\left(\epsilon_i \geq x|\mathrm{W}_i\right)=S_\epsilon\left(x\right) \quad \text{and} \quad S_\epsilon\left(\cdot\right) \quad \text{is} \quad \text{unknown.}$$

,

To make inference about ***β***, Nan et al. (2006) proposed IPW linear rank estimators for case-cohort studies under independent $\mathbb{B}$sampling. Here, we propose IPW linear rank estimators for NCC studies to further incorporate $\mathbb{F}$sampling. For simplicity, we focus on the Gehan-type estimator and consider $\widehat{\beta}$ as the minimizer of the IPW weighted objective function

$$\widehat{\mathscr{G}}\left(\beta\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\widehat{w}_i\widehat{w}_j|e_i\left(\beta\right)-e_j\left(\beta\right)|_{-}$$

where $e_i$(β) = log($X_i$) − β′W$_i$ and |$x$|$_-$ = |$x$|$I$($x$ ≤ 0). Due to the non-smoothness of $\widehat{\mathscr{G}}\left(\beta\right)$ in ***β***, it is difficult to estimate the distribution of $\widehat{\beta}$ explicitly. For standard cohort studies, Jin et al. (2001) proposed resampling procedures to approximate its distribution. Here, we extend their approach to accommodate $\mathbb{F}$sampling. Specifically, we propose to perturb $\widehat{\mathscr{G}}\left(\cdot\right)$ as

$$\widehat{\mathscr{G}}^{*}\left(\beta\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}\delta_i\widehat{w}_i^{*}\widehat{w}_j^{*}|e_i\left(\beta\right)-e_j\left(\beta\right)|_{-}$$

and obtain $\widehat{\beta}^{*}=\text{argmin}_{\beta}\widehat{\mathscr{G}}^{*}\beta$. Under the same regularity conditions given in Jin et al. (2001) and Nan et al. (2006), one may show that $n^{\frac{1}{2}}\left(\widehat{\beta}^{*}-\widehat{\beta}\right)$ conditional on the observed data $\mathscr{F}$ can be used to approximate the unconditional distribution of $n^{\frac{1}{2}}\left(\widehat{\beta}-\beta\right)$.

To estimate the survival function, $S_{\mathrm{W}_0}\left(t_0\right)=S_\epsilon\left\{\log\left(t_0\right)-\beta^{'}\mathrm{W}_0\right\}$, one may first estimate $S_\epsilon$($x$) as $\widehat{S}_\epsilon\left(x\right)=\exp\left\{-\widehat{\Lambda}_\epsilon\left(x\right)\right\}$, where

$$\widehat{\Lambda}_\epsilon\left(x\right)=\sum_{i=0}^{n}\int_0^x\frac{\delta_i\widehat{w}_i dI\left(\widehat{e}_i\leq u\right)}{\sum_{j=1}^{n}\widehat{w}_j I\left(\widehat{e}_j\geq u\right)},$$

and $\widehat{e}_i=\log\left(X_i\right)-\widehat{\beta}^{'}\mathrm{W}_i$. A plug-in estimator for S$_{\mathbf{W}0}$($t_0$) can then be obtained as $\widehat{S}_{\mathrm{W}_0}\left(t_0\right)=\widehat{S}_\epsilon\left\{\log\left(t_0\right)-\widehat{\beta}^{'}\mathrm{W}_0\right\}$. To construct interval estimators for S$_{\mathbf{W}0}$($t_0$), we extend the resampling procedure considered in Park and Wei (2003) for the AFT model under standard cohort studies and perturb $\widehat{\Lambda}_\epsilon\left(x\right)$ as

$$\widehat{\Lambda}_\epsilon^{*}\left(x\right)=\sum_{i=0}^{n}\int_0^x\frac{\delta_i^{*}\widehat{w}_i^{*}dI\left(\widehat{e}_i^{*}\leq u\right)}{\sum_{j=1}^{n}\widehat{w}_j^{*}I\left(\widehat{e}_j^{*}\geq u\right)},$$

where $\hat{e}_i^* = \log(X_i) - \mathrm{W}_i'\hat{\beta}^*$. The distribution of $n^{\frac{1}{2}}\left\{\hat{S}_{\mathrm{W}_0}(t_0) - S_{\mathrm{W}_0}(t_0)\right\}$ can then be approximated by the conditional distribution of $n^{\frac{1}{2}}\left\{\hat{S}^*_{\mathrm{W}_0}(t_0) - \hat{S}_{\mathrm{W}_0}(t_0)\right\}$ given the observed data, where $\hat{S}^*_{\mathrm{W}_0}(t_0) = \exp\left[-\hat{\Lambda}^*_\epsilon\left\{\log(t_0) - \mathrm{W}'\hat{\beta}^*\right\}\right]$. Pointwise and simultaneous CIs can be constructed for $S_{\mathbf{W}_0}(t_0)$ based on realizations of $\hat{S}^*_{\mathrm{W}_0}(t_0)$ as described in section 3·1. The justification for the resampling method for these IPW estimators under the AFT model, involving extending the arguments as given in Appendix A and C to incorporate U-processes, warrants further research.

## 4. NUMERICAL STUDIES

To examine the finite sample performance of our proposed procedures, we conducted simulation studies to assess (i) how well the resulting SE estimates approximate the true sampling standard errors and (ii) the empirical coverage probabilities of the resulting CIs. Data are generated as follows. Two models were considered to generate the survival time $T$:

$$\text{(I)} \quad \text{Cox} \quad \text{model:} \quad S_{W_0}(t) = \exp\left\{-e^{-3}t^2\exp(Y_1 + .5Y_2)\right\}, \quad (4.3)$$

$$\text{and} \quad \text{(II)} \quad \text{AFT} \quad \text{model:} \quad S_{\mathrm{W}_0}(t) = S_\epsilon\left\{\log(t) - .5Y_1 - .5Y_2\ )\right\}, \quad (4.4)$$

where $S_\varepsilon(t) = 1 - \Phi(x - 2.6)$ and $\Phi(\cdot)$ is the cumulative distribution function of a standard normal. We generate $\mathbf{Y} = (Y_1, .., Y_p)'$ from a multivariate normal with mean zero, $\mathrm{var}(Y_l) = \sigma_{ll}$ and $\mathrm{cov}(Y_l, Y_k) = \rho$. We let $\sigma_{ll} = 1.25$ and $\rho = 0.25$ for the Cox model and $\sigma_{11} = 1$, $\sigma_{22} = 2$ and $\rho = 1$ for the AFT model.

The censoring $C$ was generated as the minimum of $\mu_c(Y_1 + Y_2)$ + Gamma(2, 5)$e^{-.5Y1-.5Y2}$ and Uniform(.5,2), where $\mu_c(y) = .1$ for the Cox model and $\mu_c(y) = e^{-.2y-1}$ for the AFT model. This yielded an event rate of about 5%. We considered $n = 5000$ or 10000, and $m = 1$ or 3. We examined how resampling procedures perform when controls are sampled randomly from the risk set of cases (a) without additional matching; and (b) subject to the matching constraint of $|\mathbf{Z}_i - \mathbf{Z}_j| \le \boldsymbol{a}_0$, where $\mathbf{Z}_i = (Z_{i1}, Z_{i2})'$, $Z_{i1} = I\{\Phi(Y_{i1} + \varepsilon_{i1}) > .5\}$, $Z_{i2}$ is the closest integer to $5\Phi(Y_{i2} + \varepsilon_{i2})$, where $a_0 = (0, 1)'$, $\varepsilon_{i1}$ and $\varepsilon_{i2}$ are independent $N(0, 1)$ variables. For each configuration, 1000 datasets were simulated to summarize results of empirical estimates and $B_0 = 500$ perturbations were used for each dataset and $I_{il}$ was generated from the exponential distribution with rate 1. We note that the results are not sensitive to the choice of the distribution of $I_{il}$ and $B_0 = 500$ is generally sufficient to approximate the SEs well. For comparison, we also provide the naive resampling estimates which correspond to bootstrapping the NCC data with $\hat{w}_j$ treated as known and correspond to the robust variance.

We first summarize results for the Cox model without matching. For this setting, we also compare to the standard conditional logistic regression approach as proposed in Goldstein and Langholz (1992); Borgan et al. (1995). In the Supplementary Table 1, we present the bias, SE estimate of $\boldsymbol{\beta}$ as well as the coverage probability of the 95% CIs. All estimators have negligible bias and the resampling based CIs have proper coverage probabilities. Since the magnitude of $\boldsymbol{\beta}$ is large in this setting, the conditional logistic regression estimators are less efficient compared to the IPW estimators. In the absence of matching, the naive SE estimators for $\boldsymbol{\beta}$ without accounting for the correlation between $V_i$'s are not substantially different from the resampling based estimators. Results for $S^{\mathbf{W}0}(t_0)$ with $\mathbf{W}_0 = (0, 0)'$ and $(1, 1)'$ for $t_0 \in \mathscr{T} = [.3, 1.5]$ are shown in Figure 1. The resampling procedures yield SE estimates close to empirical SE (empirical SEs). On the other hand, the naive estimates could over estimate the variance by 30% for some $t_0$'s. The efficiency loss for the conditional logistic regression based estimators compared to the IPW estimators for $S^{\mathbf{W}0}\ (t_0)$ is quite substantial in this setting as shown in Figure 1(c). The resampling based pointwise CIs have empirical coverage probability close to the nominal level. The empirical coverage probability of simultaneous CIs are .933, .963, .941, .960 for $\{S_{(0,0)'}\ (t_0),\ t_0 \varepsilon \mathscr{T}$ .960, .964, .956, .971 for $\{S_{(1,1)'}\ (t_0),\ t_0 \varepsilon \mathscr{T}$ when $(n, m)$ = (5000, 1), (5000, 3), (10000, 1), and (10000, 3), respectively. In the presence of additional matching with **Z**, the resampling based SE estimates and CIs remain to perform well as shown in the Supplementary Figure 1. The efficiency loss for using naive variance estimates is more apparent in this matched setting especially with $m = 1$, which again signifies the importance of valid inference procedure that accounts for correlations among $V_i$'s.

For the AFT model, we only present results in the presence of matching. The results for making inference about $\beta_0 = (.5, .5)'$ are summarized in Table 1. The perturbation procedure provides SE estimates close to the empirical SEs and CIs with empirical coverage probability close to the nominal level of 95%. Results for the baseline survival curve $\{\mathrm{S}_{(0,0)'}\ (t_0) = 1 - \Phi\{\log(t) - 2.6),\ t_0 \varepsilon \mathscr{T} = [.8, 1.9]\}$ are shown in Figure 2, where $\mathrm{S}_{(0,0)'}\ (t_0)\ \varepsilon$ [.975, .997] when $t_0 \varepsilon \mathscr{T}$ The estimated SEs are again close to the empirical SEs and the pointwise 95% CIs have coverage level near the nominal level across all $t_0 \varepsilon \mathscr{T}$ For $\{\mathrm{S}_{\mathbf{0}}(t_0),\ t_0 \varepsilon \mathscr{T}$, empirical coverage probabilities of simultaneous CIs are .964, .953, .951, .957, when $(n, m)$ = (5000, 1), (5000, 3), (10000, 1), and (10000, 3), respectively. As also shown in Figure 2, ignoring the correlation among $V_i$'s would result in 20% to 40% in over-estimation of the true variance when $m = 1$ and 0% to 20% when $m = 3$.

## 5. EXAMPLE

We illustrate our proposal by an example of predicting cardiovascular disease (CVD) risk. Risk assessment plays a central role of CVD prevention. The Framingham Risk Score (FR-score) remains a widely used clinical tool for guiding the delivery of preventive cardiology of the past 30 years (D'Agostino et al., 2008; Wilson et al., 1998). However, the FR-score only has a moderate accuracy in predicting the disease occurrence. In recent years, contemporary biomarkers such as C-reactive protein (CRP) have been sought as candidates for improving the CVD risk assessment (Ridker et al., 2002, 2007). Furthermore, recent genome-wide association studies have identified genetic markers associated with CVD risk (Arnett et al., 2007). Genetic profiles, combined with novel biomarkers and modifiable environmental risk factors, may hold the key for improved prediction of disease and treatment outcomes for individuals.

One major challenge in developing such new risk models is that obtaining new markers for the entire cohort may be infeasible and/or not cost-effective. It is thus interesting to examine the effectiveness of the NCC design compared to the standard full cohort design. We illustrate this by ascertaining subject level CVD risks using a NCC sub-cohort of the Framingham Offspring Study (FOS). The FOS, initiated in 1971, enrolled a cohort of 5,124 men and women with age ranging from 5 to 70 years and followed the subjects prospectively. Here, the outcome of interest is time from the second exam date to the first major CVD event as defined previously (Lloyd-Jones et al., 2004). We are particularly interested in the short-term predictive capacity of CRP and thus censor all observations at 11 years. Since baseline risks and the effects of predictors may vary by gender, for illustration we only present the results for men. To construct a genetic risk score, we select 14 single-nucleotide polymorphisms (SNPs) that were previously reported as top 50 SNPs with p-value $< 10^{-4}$ by the Framingham Heart Study (FHS) 100k project (Larson et al., 2007) and are available in our data. A genetic score is obtained by fitting a Cox regression model with these 14 SNP using male subjects in the original cohort of the FHS. The coefficients from the fitted model are then used as weights to construct genetic scores for the FOS cohort. We consider here 1260 FOS male participants who were free of CVD and had complete information on the SNPs and CRP at the second exam, among which 109 participants had events during follow up.

To construct an NCC subcohort, for each of the 109 cases, we randomly select 3 controls, when available, from the risk set of the cases, matching on gender-standardized FR-scores (± 1 standard deviation), and age (± 15 years). We report the results by averaging estimates over 200 sampled NCC datasets. We use a multivariate Cox regression model to specify the relation between the failure time and Framingham risk score, CRP concentration (in log scale) and genetic score. The fitted coefficients from a NCC sam- pling with average sample size of 524 is comparable with those estimated from the full cohort with 2688 subjects (Table 2). Both the FR-score and CRP are predictive of CVD risks, but not the genetic score. This is consistent with the recent findings given in Paynter et al. (2010), where the hazard ratio was estimated as 1.0 for a composite gene score in a similar model for women.

In Figure 3, we present the estimated survival functions for individuals whose genetic scores are both at the median value, and FR-score and CRP levels are at the top 25% and 75% in the population respectively. The absolute risk functions estimated from much smaller NCC samples agree quite well with those estimated from the full cohorts. Furthermore, comparing the CIs of survival functions of individuals with the two risk profiles, it is apparent that individuals with high FR-score and CRP levels have significantly higher risk of developing CVD events compared to those with low FR-score and CRP levels over the 10-year period.

## 6. REMARKS

The NCC design provides a cost-effective sampling procedure for biomarker research. However the complex inference procedure often hampers its adoption in practice. Modified bootstrap procedures proposed to account for finite population sampling arising from survey sampling designs are not directly applicable to the NCC setting where the sampling is performed repeatedly over the risk sets. We propose a general yet simple perturbation-based resampling procedure for approximating the distribution of IPW estimators un- der NCC studies. The key to the proposed procedure is to perturb the IPW weights properly to recover the correlation structure among the sampling indicators of being selected into the NCC studies. The proposed approach takes advantage of the equivalence in the distribution between two IPW estimators constructed under $\mathbb{F}$sampling and $\mathbb{B}$sampling. Such an equivalence allows us to mimic the correlation structure of the sampling indicators indirectly by perturbing the sampling probabilities along with the sampling indi- cators, and the resulting estimators would be both valid and efficient regardless whether $\mathbb{B}$ or $\mathbb{F}$ sampling is used for the study. Since the perturbation for the IPW weights is not specific to the underlying survival model, the proposed procedures could be easily adapted to most survival models and thus may resolve a bottleneck in making inference with data from NCC studies.

One important advantage of the standard NCC analysis based on the conditional logistic regression is its potential in handling issues related to biomarker analyses such as effects of analytic batch and storage duration. When such effects are suspected, case and control sets can be sampled with such matching factors in a NCC design. The conditional models preserve the matching and therefore removes the potential bias, whereas the IPW approaches, derived by breaking the matching, may lead to biased estimators. Hence, when the Cox model is assumed, it would be important to conduct sensitivity analysis comparing estimators derived from different procedures. When batch effect is suspected, one may consider accounting for such effects via random effects modeling. The resampling approach we proposed can then be adopted to facilitate the inference procedure under such more complex models.

## Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

## Acknowledgments

The Framingham Heart Study and the Framingham SHARe project are conducted and supported by the National Heart, Lung, and Blood Institute (NHLBI) in collaboration with Boston University. The Framingham SHARe data used for the analyses described in this manuscript were obtained through dbGaP (access number: phs000007.v3.p2). This manuscript was not prepared in collaboration with investigators of the Framingham Heart Study and does not necessarily reflect the opinions or views of the Framingham Heart Study, Boston University, or the NHLBI. This research was in part supported by grants P01 CA053996, R01 GM085047, U01 CA086368, U54 LM008748 and R01 GM079330.

## APPENDIX

## A. JUSTIFICATION FOR THE RESAMPLING METHOD

In this section, we provide justification for the resampling method when $\widehat{\theta}_{t_0}=\text{argmin}_\theta \widehat{w}_i L_{t_0}(\theta;\text{D}_\mathbf{i})$ for a fixed $t_0$ and prove the following theorem. For simplicity, we focus on the setting without matching.

**Theorem 1.** *Assume the following regularity conditions:*

*(C1)* $E\{L_{t0}(\theta_{t0};\mathbf{D}_i)\}$ *has a unique minimum at* $\theta_{t0}$ *with continuous and positive definite* $\mathbb{A}^{t0}(\theta_{t0})$.

*(C2) The class of functions indexed by* $\theta$, $\{L_{t_0}(\theta;\mathbf{D}) \mid \theta \in \Omega^{t^0}\}$, is Glivenko-Cantelli (Kosorok, 2008) and has total variation bounded by a constant, where $\Omega_{t_0}$ *is the compact parameter space containing* $\theta_{t0}$.

*(C3) There exists a "quasi-derivative" function* $\text{U}_{t_0}(\theta;\mathbf{D})$ *for* $L_{t_0}(\theta;\mathbf{D})$ *such that for any positive sequence* $\delta_n \to 0$, *(a)* $E\left\{\text{U}_{t_0}^{\otimes 2}(\theta_{t_0};\text{D}_i)\right\}$ *is a positive definite matrix; (b)* $E\{L_{t0}(\theta;\mathbf{D}_i) - L_{t_0}(\theta_{t_0};\mathbf{D}_i)\} - \text{U}_{t0}(\theta_{t0};\text{D}i)(\theta - \theta_{t0})\} = ½(\theta - \theta_{t0})^T \mathbb{A}_{t0}(\theta - \theta_{t0}) + o(\|\theta - \theta_{t0}\|^2)$, where $\|\theta - \theta_{t0}\| \le \delta_n$;*(c)*

$$n^{-1}\sum_{i=1}^{n}\widehat{w}_i\left\{L_{t_0}(\theta_1;\text{D}_i) - L_{t_0}(\theta_2;\text{D}_i) - \text{U}_{t_0}(\theta_2;\text{D}_i)(\theta_1-\theta_2)\right\}$$

$$=\frac{1}{2}(\theta_1-\theta_2)^T\mathbb{A}_{t_0}(\theta_1-\theta_2)+_o\left(\|\theta_1-\theta_2\|^2+n^{-\frac{1}{2}}\|\theta_1-\theta_2\|\right)$$, almost surely, uniformly in $\|\theta_1 - \theta_{t0}\| \le \delta_n$, $\|\theta_2 - \theta_{t0}\|_2 \le \delta_\text{n}$.

*Then* $\widehat{\theta}_{t_0}^* \to \theta_{t_0}$ *in probability and* $\widehat{W}_{t_0}^*=n^{\frac{1}{2}}\left(\widehat{\theta}_{t_0}^*-\widehat{\theta}_{t_0}\right)$ *conditional on* $\mathscr{F}$ *converges in distribution to* $N(0, \Sigma_{t0})$, *the limiting distribution of* $\widehat{W}_{t_0}$.

*Remark 1:* Conditions C1-C3 are parallel to the conditions required in Proposition A1-A3 in Jin et al. (2001) and guarantee that $\widehat{\theta}_{t_0}$ is a consistent estimator of $\theta_{t0}$ and $\widehat{W}_{t_0} \to N\left(0, \sum_{t_0}\right)$ in distribution.

*Remark 2:* Although these conditions often need to be verified on a case by case basis, smooth functionals with bounded second derivate functions, convexity or quasi-derivative functions with total variation bounded by a constant would often lead to objective functions that satisfy these conditions.

*Remark 3:* The smoothness conditions are required for the expected loss functions while the loss function itself does not have to be very smooth.

*Proof.* Let ℙ and ℙ denote the probability measure generated by F and F × I, respectively. We first show that $\widehat{\theta}^*_{t_0} \rightarrow \ _* \theta_{t_0}$. To this end, we first note that by a law of large numbers and the monotonicity of $\exp\left\{-\widehat{\Lambda}^*_{\mathrm{marg}}(t)\right\} - \widehat{G}(t) \rightarrow \ _*$ uniformly in *t*. It follows that $\max|\widehat{p}_{0_j}/\widehat{p}^*_{0_j} - 1| =_{O \ *}$ (1) This, together with condition (C2), implies that

$|$

$+|\widehat{L}_{t_0}(\theta)$

$- E\left\{L_{t_0}(\theta_{t_0}, \mathrm{D})\right\}| +_{O \ *} (1)$, which uniformly converges to 0. Then under condition C1, $E\{L_{t0}(\theta_{t0}, \mathbf{D})\}$ has a unique minimum at $\theta_{t0}$ and hence from Newey and McFadden (1994), $\widehat{\theta}^*_{t_0} \rightarrow \ _* \theta_{t_0}$.

We next derive the distribution for $\widehat{\theta}^*_{t_0}$. Consider $\theta = \theta_t + n^{-1/2}$u. Condition (C3) implies that

$$\frac{\widehat{L}_{t_0}\left(\theta_{t_0} + n^{-\frac{1}{2}}\mathrm{u}\right) - \widehat{L}_{t_0}(\theta_{t_0}) - \mathrm{u}'\widehat{W}_{\mathrm{U}_{t_0}} - \frac{1}{2}n^{-1}\mathrm{u}'\mathbb{A}_{t_0}\mathrm{u}}{\|n^{-\frac{1}{2}}\mathrm{u}\|} = O\mathbb{P}(1)$$

uniformly in U. It then follows from similar arguments as those used for the proof of the multiplier central limit theorem (CLT) (Kosorok, 2008, Theorem 10.1) that

$$\frac{\widehat{L}^*_{t_0}\left(\theta_{t_0} + n^{-\frac{1}{2}}\mathrm{u}\right) - \widehat{L}^*_{t_0}(\theta_{t_0}) - \mathrm{u}'\widehat{W}^*_{\mathrm{U}_{t_0}} - \frac{1}{2}n^{-1}\mathrm{u}'\mathbb{A}_{t_0}\mathrm{u}}{\|n^{-\frac{1}{2}}\mathrm{u}\|} = O\mathbb{P}*(1)$$

where $\widehat{W}^*_{\mathrm{U}_{t_0}} = n^{-\frac{1}{2}}\sum_{i=1}^{n}\widehat{w}^*_i \mathrm{U}_{t_0}(\theta_{t_0}; \mathrm{D}_i)$. It follows that $\widehat{W}^*_{t_0} = \mathbb{A}^*_{t_0}\left\{\widehat{W}^{-1}_{\mathrm{U}_{t_0}} - \widehat{W}_{\mathrm{U}_{t_0}}\right\} +_{O_p}(1)$. We next show that $\widehat{W}^*_{\mathrm{U}_{t_0}} - \widehat{W}_{\mathrm{U}_{t_0}}$ conditional on $\mathscr{F} \rightarrow N$ (0, $C^{Ut0}$ in distribution. To this end, we write $\widehat{W}^*_{\mathrm{U}_{t_0}} - \widehat{W}_{\mathrm{U}_{t_0}} = \widehat{W}^*_{1\mathrm{U}_{t_0}} + \widehat{W}^*_{2\mathrm{U}_{t_0}}$, where

$$\widehat{W}^*_{1\mathrm{U}_{t_0}} = n^{-\frac{1}{2}}\sum_{j=1}^{n}\delta_j(I_{jj} - 1)\,\mathrm{U}_{t_0}(\mathrm{D}_j) + n^{-\frac{1}{2}}\sum_{j=1}^{n}\frac{V^*_{0j} - V_{0j}}{\widehat{p}_{0j}}(1 - \delta_j)\,\mathrm{U}_{t_0}(\mathrm{D}_j), \quad \text{(A.1)}$$

$$\widehat{W}^*_{2\mathrm{U}_{t_0}} = n^{-\frac{1}{2}}\sum_{j=1}^{n}\left(\frac{V^*_{0j}}{\widehat{p}^*_{0j}} - \frac{V^*_{0j}}{\widehat{p}_{0j}}\right)(1 - \delta_j)\,\mathrm{U}_{t0}(\mathrm{D}_j) \quad \text{(A.2)}$$

First, we note that since $\{I_{ij}\}$'s are independent and $E\left(I^2_{ij}\right) = 2, \quad \mathrm{var}\left(V^*_{0j}|\mathscr{F}\right) = \Pi_{i:X_i \le X_j, \delta_i = 1}(1 - 2V_{0ij} + 2V_{0ij}) - (1 - V_{0j}) = V_{0j}$. Moreover, it is straightforward to see that $\{I_{jj}\delta_j$, $\left\{I_{jj}\delta_j, V^*_{0j}(1 - \delta_j), j = 1, ..., n\right\}$ $(1 - \delta_j)$,$_j$ = 1, ..., *n*} are independent of each other conditional of $\mathscr{F}$. Thus,

$\text{var}\left(\widehat{W}^{*}_{1\mathrm{U}_{t_0}}|\mathscr{F}\right)=n^{-1}\sum_{i=1}^{n}\left\{V_j/\widehat{p}_j^2\right\}\mathrm{U}_{t_0}(\mathrm{D}_j)^{\otimes 2}\to E\left\{\mathrm{U}_{t_0}(\mathrm{D}_j)^{\otimes 2}/p_j\right\}$ in probability. It then follows from a Linderberg CLT that $\widehat{W}^{*}_{1\mathrm{U}_{t_0}}$ given $\mathscr{F}\to\mathscr{F}\to N\left(0,E\left\{\mathrm{U}_{t_0}(D_j)^{\otimes 2}/p_j\right\}\right)$ in distribution.

Next, by a taylor series expansion, $\widehat{W}^{*}_{2\mathrm{U}_{t_0}}$ is asymptotically equivalent to

$$-n^{-\frac{1}{2}}\sum_{j=1}^{n}\frac{V^{*}_{0j}\left(1-\delta_j\right)}{\widehat{p}_{0j}\widehat{p}^{*}_{oj}}\widehat{G}\left(X_j\right)\left\{\widehat{\Lambda}^{*}_{\text{marg}}\left(X_j\right)-\widehat{\Lambda}_{\text{marg}}\left(X_j\right)\right\}\mathrm{U}_{t_0}\left(\mathrm{D}_j\right)\approx\int n^{\frac{1}{2}}\left\{\widehat{\Lambda}^{*}_{\text{marg}}\left(t\right)-\widehat{\Lambda}_{\text{marg}}\left(t\right)\right\}d\widehat{\eta}^{*}_{\mathrm{U}_{t_0\left(\mathrm{D}_j\right)}}\left(t\right)$$

$$\begin{aligned}&\widehat{\eta}^{*}_{\mathrm{U}_{t_0}}\left(t\right)\\&\quad=n^{-1}\sum_{j=1}^{n}V^{*}_{0j}\left(1-\delta_j\right)\\&\quad/\widehat{p}^{2}_{0j}\widehat{G}\left(X_j\right)I\left(X_j\geq t\right)\mathrm{U}_{t_0}\left(D_j\right)\quad\text{and}\quad\widehat{\Lambda}_{\text{marg}}\left(t\right)\\&\quad=\sum_{i=1}^{n}N_i\left(t\right)m\\&\quad/\|\mathscr{R}_i\|\\&\quad=\sum_{i=1}^{n}N_i\left(t\right)\sum_{l=1}^{n}I\left(X_l\geq X_i\right)V_{0il}\\&\quad/\|\mathscr{R}_i\|\quad\text{with}\quad\widehat{\Lambda}_{\text{marg}}\left(t\right)\end{aligned}$$

where $-\log\widehat{G}\left(t\right)=_{O_p}\left(n^{-\frac{1}{2}}\right)$. It follows from a uniform law of large numbers (LLN) (Pollard, 1990) that $\sup_t|\widehat{\eta}^{*}_{\mathrm{U}_{t_0}}\left(t\right)-\eta_{\mathrm{U}_{t_0}}\left(t\right)|\to 0$, in probability. We next write

$$n^{\frac{1}{2}}\left\{\widehat{\Lambda}^{*}_{\text{marg}}\left(t\right)-\widehat{\Lambda}_{marg}\left(t\right)\right\}=n^{\frac{1}{2}}\sum_{i=1}^{n}\widehat{U}^{*}_{i}N_i\left(t\right),\quad\widehat{U}^{*}_{i}=\frac{\sum\limits_{l=1}^{n}I\left(X_l\geq X_i\right)V_{0il}\left(I_{il}-1\right)}{\|\mathscr{R}\|}$$

and then use an integration by parts to obtain

$$\widehat{W}^{*}_{2\mathrm{U}_{t_0}}\approx -n^{\frac{1}{2}}\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}\left(X_i\right)\delta_i\widehat{U}^{*}_{i}\quad\text{(A.3)}$$

Since $\left\{\eta_{\mathrm{U}_{t_0}}\left(X_i\right)\widehat{U}^{*}_{i}\delta_i,i=1,\dots,n\right\}$ are independent random variables conditional on $\mathscr{F}$

$$\text{var}\left(\widehat{W}^{*}_{2\mathrm{U}_{t_0}}\quad|\quad\mathscr{F}\right)=n\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}\left(X_i\right)^{\otimes 2}\delta_i\text{var}\left(\widehat{U}^{*}_{i}\quad|\quad\mathscr{F}\right)=n\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}\left(X_i\right)^{\otimes 2}\delta_i\frac{\sum\limits_{l=1}^{n}I\left(X_l\geq X_i\right)V_{0il}}{\|\mathscr{R}\|^2}$$

It follows that $\text{var}\left(\widehat{W}_{2\mathrm{U}_{t_0}}|\mathscr{F}\right)=n\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}\left(X_i\right)^{\otimes 2}\delta_i\sum_{l=1}^{n}I\left(X_l\geq X_i\right)V_{0il}/\|\mathscr{R}\|^2$ By a LLN and the fact that $P(V_{0il} = 1 \mid \mathscr{F}) = m\|\mathscr{R}_{\mathrm{i}}\|$, we have

$$n\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}(X_i)^{\otimes 2}\delta_i\frac{\sum_{l=1}^{n}I(X_l\geq X_i)V_{0il}}{\|\mathscr{R}\|^2}\to mE\left\{\delta_i\frac{\eta_{\mathrm{U}_{t_0}}(X_i)^{\otimes 2}}{\pi(X_i)^2}\right\}=m\int\frac{\eta_{\mathrm{U}_{t_0}}(t)^{\otimes 2}}{\pi(t)}d\Lambda_{\mathrm{marg}}(t),\quad \text{in}\quad \text{probability.}$$

in probability.

Thus, by a Linderberg CLT, $\mathscr{F}$ given converges in distribution to $N\left(0, m\int\eta_{\mathrm{U}_{t_0}}(t)^{\otimes 2}\pi(t)^{-1}d\Lambda_{\mathrm{marg}}(t)\right)$.

To obtain the correlation between $\widehat{W}^{*}_{1\mathrm{U}_{t_0}}$ and $\widehat{W}^{*}_{2\mathrm{U}_{t_0}}$ we note that since $\{I_{ij}\}$ are independent of $\{I_{ij}, i \neq j\}$ and $V^{*}_{0j}$ is independent of $I_{ij'}$ when $j \neq j'$,

$$\mathrm{cov}\left(\widehat{W}^{*}_{1\mathrm{U}_{t_0}}, \widehat{W}^{*}_{2\mathrm{U}_{t_0}}\quad|\quad\mathscr{F}\right)\approx-\sum_{j=1}^{n}\frac{(1-\delta_j)\,\mathrm{U}_{t_0}(\mathrm{D}_j)}{\widehat{p}_{0j}}\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}(X_i)'\,\delta_i\frac{I(X_j\geq X_i)}{\|\mathscr{R}\|}V_{0ij}\mathrm{cov}\left(V^{*}_{0j}, I_{ij}|\mathscr{F}\right).$$

Now,

$\mathrm{cov}\left(V^{*}_{0j}, I_{ij}|\mathscr{F}\right)=-E\left\{\Pi_{i':j\epsilon\mathscr{R}_{i'}}\left(1-\delta_{i'}V_{0i'j}I_{i'j}\right)I_{ij}\right\}+(1-V_{0j})=V_{0ij}\Pi_{i':j\epsilon\mathscr{R}_{i'}}\left(1-\delta_{i'}V_{0i'j}\right)$ . Therefore,

$$\mathrm{cov}\left(\widehat{W}^{*}_{1U_{t_0}}, \widehat{W}^{*}_{2\mathrm{U}_{t_0}}\quad|\quad\mathscr{F}\right)=-\sum_{j=1}^{n}\frac{(1-\delta_j)\,\mathrm{U}_{t_0}(D_j}{}\,)\widehat{p}_{0j}\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}(X_i)'\,\delta_i\frac{I(X_j\geq X_i)}{\|\mathscr{R}\|}V_{0ij}\prod_{i':j\epsilon\mathscr{R}_{i'}}(1-\delta_{i'}V_{0i'j})$$

It is not difficult to see that

$E\left\{V_{0ij}\Pi_{i':j\epsilon\mathscr{R}_{i'}}\left(1-\delta_{i'}V_{0i'j}\right)|\mathscr{D}\right\}=E(V_{0ij}|\mathscr{D})\,\Pi_{i\neq i'}\left\{1-E\left(V_{0i'j}|\mathscr{D}\right)\right\}\approx m(1-\widehat{p}_{0j})/\|\mathscr{R}_i\|$

. It follows the $E\left\{\mathrm{cov}\left(\widehat{W}^{*}_{1\mathrm{U}_{t_0}}, \widehat{W}^{*}_{2\mathrm{U}_{t_0}}|\mathscr{F}\right)|\mathscr{D}\right\}$

$$\approx -m\sum_{j=1}^{n}\frac{(1-\delta_j)\,\mathrm{U}_{t_0}(\mathrm{D}_j)}{\widehat{p}_{0j}}\sum_{i=1}^{n}\eta_{\mathrm{U}_{t_0}}(X_i)\,\delta_i\frac{I(X_j\geq X_i)}{\|\mathscr{R}\|}\frac{(1-\widehat{p}_{0j})}{\|\mathscr{R}\|}\to -m\int\eta_{\mathrm{U}_{t_0}}(t)^{\otimes 2}\pi(t)^{-1}d\Lambda(t),$$

In probability. Putting together the variance and covariance of $\widehat{W}^{*}_{1\mathrm{U}_{t_0}}$ and $\widehat{W}^{*}_{2\mathrm{U}_{t_0}}$ and applying the Linder-berg CLT again, we have $\widehat{W}^{*}_{\mathrm{U}_{t_0}}-\widehat{W}_{\mathrm{U}_{t_0}}$ given $\mathscr{F}$ converges in distribution to $N(0, \mathbb{C}_{\mathrm{U}t0})$. When $\theta_{t0}$ changes over $t_0$ or when the objective function for obtaining $\widehat{\theta}_{t_0}$ is not a simple IPW weighted sum of independent and identically distributed terms, one may need justify why $\widehat{W}^{*}_{t_0}$ can be used for approximating the distribution of $\widehat{W}_{t_0}$ as a process in $t_0$ on a case by case basis. For illustration, we outline the justification for the Cox model in Appendix C.

## B. VARIANCE APPROXIMATION IN THE PRESENCE OF MATCHING

In the presence of matching, the correlation between $V^{\dagger}_{0i}$ and $V^{\dagger}_{0j}$ is

$$\widehat{\rho}_{ij}^{\dagger}=\prod_{X_k\le\min(X_i,X_j),\delta_k=1}\frac{1-\frac{m\{I(|\mathrm{Z}_k-\mathrm{Z}_i|\le a_0)+I(|\mathrm{Z}_k-\mathrm{Z}_j|\le \mathrm{a}_0)\}}{\|\mathscr{R}_k^{\dagger}\|}+\frac{m(m-1)I(|\mathrm{Z}_k-\mathrm{Z}_i|\le a_0,|\mathrm{Z}_k-\mathrm{Z}_j|\le \mathrm{a}_0}{\left\{\|\mathscr{R}_k^{\dagger}\|\right\}\left\{\|\mathscr{R}_k^{\dagger}\|-1\right\}}}{\left\{1-\frac{mI(|\mathrm{Z}_k-\mathrm{Z}_i|\le \mathrm{a}_0)}{\|\mathscr{R}_k^{\dagger}\|}\right\}\left\{1-\frac{mI(|\mathrm{Z}_k-\mathrm{Z}_j|\le \mathrm{a}_0)}{\|\mathscr{R}_k^{\dagger}\|}\right\}}-1.$$

Following from similar arguments as given Samuelsen (1997), we have

$$\widehat{\rho}_{ij}^{\dagger}=-mn^{-1}\int\int\frac{I(X_i\ge t,|\mathrm{Z}_i-\mathrm{z}|\le a_0)I(X_j\ge t,|\mathrm{Z}_j-\mathrm{z}|\le a_0)}{\pi(t,\mathrm{z})^2}A_{\mathrm{z}}\left(dt\right)\mathbb{F}\left(dz\right)+O_p\left(n^{-3/2}\right)$$

where $m\left(1-\widehat{p}_{0j}\right)/\|\mathscr{R}_i\|$. and $E\left\{\mathrm{cov}\left(\widehat{W}_{1\mathrm{U}_{t_0}}^{*},\widehat{W}_{2\mathrm{U}_{t_0}}^{*}|\mathscr{F}\right)|\mathscr{D}\right\}$ It follows that

$$\widehat{r}_{ij}^{\dagger}=\mathrm{cov}\left(\widehat{w}_i^{\dagger},\widehat{w}_j^{\dagger}|\mathscr{D}\right)=-n^{-1}m\int\int\frac{\eta(t,\mathrm{z};X_i,\delta_i)\eta(t,\mathrm{z};X_j,\delta_j)}{\pi(t,\mathrm{z})^2}A_{\mathrm{z}}\left(dt\right)\mathbb{F}\left(dz\right)+O_p\left(n^{-3/2}\right)$$

$$\mathbb{C}_{\mathrm{U}_{t_0}}^{\dagger}=E\left\{\frac{\mathrm{U}_{t_0}(\mathrm{D}_i)^{\otimes 2}}{p_i\dagger}\right\}-m\int\int\frac{\eta_{\mathrm{U}_{t_0}}(u,\mathrm{z})^{\otimes 2}}{\pi(u,\mathrm{z})^2}A_{\mathrm{z}}\left(du\right)\mathbb{F}\left(dz\right)\quad\text{(A.4)}$$

where $\eta\left(t,\mathrm{z};X_i,\delta_i\right)=I\left(X_i>t,|\mathrm{Z}_i-\mathrm{z}|\le a_0\right)\left(1-p_i\right)/p_i$. Then it is not difficult to show that the random vector $n^{-\frac{1}{2}}\sum_{i=1}^{n}\widehat{w}_i^{\dagger}\mathrm{U}_{t_0}\left(\mathrm{D}_i\right)=n^{-\frac{1}{2}}\sum_{i=1}^{n}\left(\widehat{w}_i^{\dagger}-1\right)\mathrm{U}_{t_0}\left(\mathrm{D}_i\right)+n^{-\frac{1}{2}}\sum_{i=1}^{n}\mathrm{U}_{t_0}\left(\mathrm{D}_i\right)$ has asymptotic variance where $\int_0^t\pi(u,\mathrm{Z})^{-1}A_{\mathrm{z}}\left(du\right)\mathbb{F}\left(d\mathrm{Z}\right)$. Since the matching only affects the correlation structure between $V_{0j}^{\dagger}$ and $V_{0j}^{\dagger}$ and $E\left(\widehat{w}_j^{\dagger}|\mathscr{D}\right)=1$, the asymptotic variance of $\widehat{W}_{t_0}$ is $\mathbb{A}_{t_0}(\theta_{t_0})^{-1}\mathbb{C}_{\mathrm{U}_{t_0}}^{\dagger}(\theta_{t_0})^{-1}$.

## C. RESAMPLING PROCEDURE JUSTIFICATION UNDER THE COX MODEL

In this section, we show that the resampling method can be used to approximate the distribution of $\widehat{S}_{\mathrm{W}_0}\left(t_0\right)\equiv\exp\left\{-\widehat{\Lambda}_0\left(t_0\right)e^{\widehat{\beta}'\mathrm{W}_0}\right\}$ under the Cox model. From Cai and Zheng (2012), $\widehat{W}_{t_0}=n\frac{1}{2}\left\{\widehat{S}_{\mathrm{W}_0}\left(t_0\right)-S_{\mathrm{W}_0}\left(t_0\right)\right\}=n^{-\frac{1}{2}}\sum_{i=1}^{n}\widehat{w}_iU_{S,t_0}\left(\mathrm{D}_i\right)+_{O_p}\left(1\right)$ which converges weakly to a zero-mean Gaussian process in $t_0$, $W_{t0}$, with covariance function

$$\begin{aligned}
\sum\nolimits_{t_0,T_0} &= E\left\{U_{S,t_0}\left(\mathrm{D}_i\right)U_{S,T_0}\left(\mathrm{D}_i\right)/p_i\right\}\\
&- m\int \eta_{t_0}\left(t\right)\eta_{T_0}\left(t\right)^{-1}d\Lambda_{\mathrm{marg}}\left(t\right), \quad \text{where} \quad U_{S,t}\left(\mathrm{D}_i\right)\\
&= -S_{W_0}\left(t\right)\left\{e^{\beta' W_0}U_{\Lambda,t}\left(\mathrm{D}_i\right)+\Lambda_0\left(t\right)e^{\beta' W_0}U_{\beta}\left(\mathrm{D}_i\right)\right\}, \eta_{t_0}\left(t\right)\\
&= E\left\{U_{S,t_0}\left(\mathrm{D}_i\right)I\left(X_i>t\right)\left(1-p_i\right)/p_i\right\}, U_{\Lambda,t}\left(\mathrm{D}_i\right)\\
&= \int_0^t (u)^{-1}\left\{dM_i\left(u\right)-\mathrm{U}_{\beta}(\mathrm{D}_i)'\Pi_1\left(u\right)\right\}, U_{\beta}\left(\mathrm{D}_i\right)\\
&= \left[\int\left\{\Pi^{(2)}\left(t\right)\Pi_0\left(t\right)-\Pi_1\left(t\right)\right\}/\Pi_0\left(t\right)dt\right]^{-1}\int\left\{\mathrm{W}_i-\Pi_1\left(t\right)/\Pi_0\left(t\right)\right\}dM_i\left(t\right), \Pi_k\left(t,\beta\right)\\
&= E\left\{I\left(X_i\geq t\right)e^{\beta'\mathrm{W}_i}\mathrm{W}_i^{\otimes k}\right\}, \Pi_k\left(t\right)\\
&= \Pi_k\left(t,\beta_0\right) \quad \text{and} M_i\left(t\right)\\
&= N_i\left(t\right)-\int_0^t I\left(X_{i\geq u}\right)e^{\beta'\mathrm{W}_i}d\Lambda_0\left(u\right) \qquad .
\end{aligned}$$

The main results are summarized in the following theorem:

**Theorem 2.** Under the proportional hazards model, we assume that **W**'s are bounded and β$_0$ *is an interior point of a compact set* Ω as well as the regularity conditions A-D required in Andersen and Gill (1982). Then the distribution of $n^{\frac{1}{2}}\left(\widehat{\beta}^*-\widehat{\beta}\right)$ given $\mathscr{F}$ and the distribution of $n^{\frac{1}{2}}\left(\widehat{\beta}-\beta_0\right)$ converge to the same limiting normal distribution. Furthermore, for $t_0 \in \mathscr{F}$ and a given **W**$_0$, $\widehat{W}^*_{t_0}=n^{\frac{1}{2}}$ ] $\left\{\widehat{S}^*_{\mathrm{W}_0}\left(t_0\right)-\widehat{S}_{\mathrm{W}_0}\left(t_0\right)\right\}$ *given* $\mathscr{F}$*converges weakly to W$_{t0}$*.

*Proof.* Let $\widehat{l}\left(\beta,\mathrm{D}_j\right)=\int\left\{\beta'\mathrm{W}_i-\log\widehat{\Pi}_0\left(t_0,\beta\right)\right\}dN_j\left(t\right)$ and write

$$\widehat{\ell}^*\left(\beta\right)=n^{-1}\sum_{i=1}^{n}\widehat{w}_j^*\widehat{\ell}\left(\beta,\mathrm{D}_j\right)-\int\left\{\log\widehat{\Pi}_0^*\left(t,\beta\right)-\log\widehat{\Pi}_0\left(t,\beta\right)\right\}d\left\{n^{-1}\sum_{i=1}^{n}\widehat{w}_j^*N_j\left(t\right)\right\}$$

We first show that $\widehat{W}^*_{\Pi_k}\left(t,\beta\right)=n^{\frac{1}{2}}\left\{\widehat{\Pi}_k^*\left(t,\beta\right)-\widehat{\Pi}_k\left(t,\beta\right)\right\}=n^{-\frac{1}{2}}\sum\nolimits_{i=1}^{n}\left(\widehat{w}_i^*-\widehat{w}_i\right)U_{\Pi_k}\left(t,\beta;\mathrm{D}_j\right)$ converges weakly to a Gaussian process in *t* and *β*, where $U_{\Pi_k}\left(t,\beta;\mathrm{D}_j\right)=I\left(X_j\geq t\right)\mathrm{W}_j^{\otimes k}e^{\beta'\mathrm{W}_j}-\Pi_k\left(t,\beta\right)$. The pointwise convergence of $\widehat{W}_{\Pi_k}\left(1,\beta\right)$ follows from the same arguments as given in Appendix A. To establish the convergence as a process, we write $\widehat{W}^*_{\Pi_k}\left(t,\beta\right)=\widehat{W}^*_{1\Pi_k}\left(t,\beta\right)\widehat{W}^*_{2\Pi_k}\left(t,\beta\right)$ and it suffices to show that both components are tight in (*t*, *β*), where $\widehat{W}^*_{1\Pi_k}\left(t,\beta\right)$ and $\widehat{W}^*_{2\Pi_k}\left(t,\beta\right)$ are obtained by replacing U$_t$(D$_j$) in (A.1) and (A.2) with $U_{\Pi k}\left(t,\beta;\mathrm{D}_j\right)$. From the multiplier CLT (Kosorok, 2008, Theorem 10.1), $\widehat{W}^*_{1\Pi_k}\left(t_0,\beta\right)$, a sum of independent random processes, converges weakly to a zero-mean Gaussian process and hence is tight. For $\widehat{W}^*_{2U_{M,t_0}}$, by a uniform LLN (Pollard, 1990), conditional on $\mathscr{F}$,

$$\begin{aligned}
&\widehat{\eta}^*_{\Pi_k,t_0,\beta}\left(u\right)\\
&\quad = n^{-1}\sum\nolimits_{j=1}^{n}V_{0j}^*\left(1-\delta_j\right)\\
&\quad /\widehat{p}_{0j}^2\widehat{G}\left(X_j\right)I\left(X_j\geq u\right)U_{\Pi_k}\left(t_0,\beta;\mathrm{D}_j\right)\rightarrow\eta_{\Pi_k,t_0\beta}\left(u\right)\\
&\quad = E\left\{I\left(X_j>u\right)\left(1-p_j\right)/p_jU_{\Pi_k}\left(t_0,\beta;\mathrm{D}_j\right)\right\}
\end{aligned}$$

in probability, uniformly in ($t_0$, *u*, *β*). Thus, the approximation given in (A.3) holds uniformly in ($t_0$, *β*) in this case. From the multiplier CLT again, we have the weak convergence and hence the tightness of $\widehat{W}^*_{2\Pi_k}(t_0,\beta)$. This, together with the uniform convergence of $n^{-1}\sum_{i=1}^n \widehat{w}_j^* N_j(t_0) \to \int_0^{t_0}\Pi_0(t,\beta_0)\,d\Lambda_0(t)$, implies that for $\beta=\widehat{\beta}+O_p\left(n^{-\frac{1}{2}}\right), \widehat{l}^*(\beta)=\widetilde{l}^*(\beta)+O_{\ *}\left(n^{-1}\right)$, and hence $n^{\frac{1}{2}}\left(\widehat{\beta}^*-\widehat{\beta}\right)=n^{\frac{1}{2}}\left(\widetilde{\beta}^*-\widehat{\beta}\right)+_{O\ *}(1)$, where $\widetilde{\beta}^*=\text{argmin}_\beta\widetilde{l}^*(\beta)$,

$$\widetilde{\ell}^*(\beta)=n^{-1}\sum_{i=1}^n\widehat{w}_j^*\widehat{\ell}(\beta,\mathrm{D}_j)-\left(\beta-\widehat{\beta}\right)'n^{-1}\sum_{i=1}^n\widehat{w}_j^*\int\left\{\mathrm{W}_j-\frac{\widehat{\Pi}_1\left(t_0,\widehat{\beta}\right)}{\widehat{\Pi}_0\left(t_0,\widehat{\beta}\right)}\right\}I\left(X_j>t\right)e^{\widehat{\beta}'\mathrm{W}_j}d\Lambda_0(t)$$

and $p^*$is the probability measure generated by both $\mathscr{F}$and *I*. This, together with similar arguments as given in section A along with the uniform convergence of $\widehat{\Pi}_k(t_0,\beta)\to\Pi_k(t_0,\beta)$, implies that conditional on $\mathscr{F}$, $n^{\frac{1}{2}}\left(\widehat{\beta}^*-\widehat{\beta}\right)=n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)\mathrm{U}_\beta(\mathrm{D}_i)+_{O\ *}(1)\to$ the limiting distribution of $n^{\frac{1}{2}}\left(\widehat{\beta}-\beta_0\right)$.

We next establish the weak convergence of $\widehat{W}^*_{t_0}$ as a process in $t_0$. By a taylor series expansion and the consistency of $\left\{\widehat{\Pi}_k(u),\widehat{\Lambda}_0(t_0),\widehat{\beta}\right\}$ as well as their perturbed counterparts,

$$\begin{aligned} n^{\frac{1}{2}}\left\{\widehat{\Lambda}_0^*(t_0)-\widehat{\Lambda}_0(t_0)\right\} &=n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)\int_0^{t_0}\frac{dN_i(u)}{\widehat{\Pi}_0^*\left(u;\widehat{\beta}\right)}-n^{-1}\sum_{i=1}^n\widehat{w}_i\int_0^{t_0}\frac{\widehat{\Pi}_0^*\left(u;\widehat{\beta}^*\right)-\widehat{\Pi}_0\left(u;\widehat{\beta}\right)}{\widehat{\Pi}_0\left(u;\widehat{\beta}\right)\widehat{\Pi}_0^*\left(u;\widehat{\beta}^*\right)}dN_i(u) \\ &\approx n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)\int_0^{t_0}\frac{dM_i(u)}{\Pi_0(u)}-\left(\widehat{\beta}^*-\widehat{\beta}\right)'\int_0^{t_0}\frac{\Pi_1(u)}{\Pi_0(u)}d\Lambda_0(u) \end{aligned}$$

and thus $n^{\frac{1}{2}}\left\{\widehat{W}_0^*(t_0)\widehat{\Lambda}_0(t_0)\right\}\approx n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)U_{\Lambda,t_0}(\mathrm{D}_i)$. It follows that

$$\begin{aligned} \widehat{W}^*_{t_0} &=-S_{\mathrm{W}_0}(t)\left[e^{\beta'\mathrm{W}_0}n^{\frac{1}{2}}\left\{\widehat{\Lambda}_0^*(t_0)-\widehat{\Lambda}_0(t_0)\right\}+\Lambda_0(t_0)\,e^{\beta'\mathrm{W}_0}n^{\frac{1}{2}}\left(\widehat{\beta}^*-\widehat{\beta}\right)\right]+0\mathbb{P}*(1) \\ &=n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)U_{S,t_0}(\mathrm{D}_i)+O\mathbb{P}*(1). \end{aligned}$$

Using the justification similar to those given above for $\widehat{W}^*_{\Pi_k}(t,\beta)$, one may obtain the finite dimensional in distribution convergence of $\widehat{W}^*_{t_0}$ to a multivariate normal. Now, for the convergence as a process, it suffices to show that the process is tight. Since $\Lambda_0(t_0)$ and $\boldsymbol{S}_{\mathbf{W}}0(t_0)$ are deterministic functions, we only need to show that $n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)U_{M,t_0}(\mathrm{D}_i)$, where $U_{M,t_0}(\mathrm{D}_i)=\int_0^{t_0}\Pi_0(u)^{-1}dM_i(u)$ is tight. To this end, we also write $n^{-\frac{1}{2}}\sum_{i=1}^n\left(\widehat{w}_i^*-\widehat{w}_i\right)U_{M,t_0}(\mathrm{D}_i)=\widehat{W}^*_{1U_{M,t_0}}+\widehat{W}^*_{2U_{M,t_0}}$, where $\widehat{W}^*_{1U_{M,t_0}}$ and $\widehat{W}^*_{2U_{M,t_0}}$ are defined based on (A.1) and (A.2) by replacing $\mathrm{U}_{t0}$ with $U_{M,t0}$. Then following the same argument as for the tightness of $\widehat{W}^*_{1\Pi_k}(t_0,\beta)$ and $\widehat{W}^*_{2\Pi_k}(t_0,\beta)$ as given above, we have the tightness of $\widehat{W}^*_{1U_{M,t_0}}$ and $\widehat{W}^*_{2U_{M,t_0}}$. This concludes the weak convergence of $\widehat{W}^*_{t_0}$ conditional on $\mathscr{F}$to $W_{t0}$.

Figure 1: Simulation results for the survival function at marker level $\mathbf{W}_0 = (0,0)'$ (thinner curves) and $\mathbf{W}_0 = (1, 1)'$ (thicker curves) under the Cox model without additional matching variables for sample sizes (i) n = 5000 (black curves) and (ii) n = 10000 (gray curves): (a) empirical (solid) and estimated (dashed) standard errors; (b) empirical coverage probabilities of the 95% confidence intervals; and (c) efficiency loss in the naive variance estimates (solid curves) and the conditional logistic regression based estimators (dot-dashed curves).

Figure 2: Simulation results for the baseline survival function $\{\mathbf{S}_{\mathbf{W}0}(t) : t \in [.8, 1.]\}$ under the AFT model with two matching variables when sample sizes (i) n = 5000 (black curves) and (ii) n = 10000 (gray curves): (a) empirical (solid) and estimated (dashed) standard errors; (b) empirical coverage probabilities of the 95% confidence intervals; and (c) efficiency loss in the naive variance estimates.

Figure 3: Estimated absolute risks (solid lines) along with their 95% pointwise (shaded region) and simultaneous (dashed lines) CIs from the FOS based on the full cohort and an NCC design with m = 3 controls, and stratified by age and framingham score. (a) absolute risk for an individual with CRP and FR-score both at 10th percentiles respectively and gene score at 50th percentile of the population; (b) with absolute risk for an individual with CRP and FR-score both at 90th percentiles respectively and gene score at 50th percentile of the population. Shown also are the estimated absolute risks from the full cohort (dotted lines).

## REFERENCES

Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. Ann. Statist. 1982:1100–1120.

Arnett D, Baird A, Barkley R, Basson C, Boerwinkle E, Ganesh S, Herrington D, Yuling H, Jaquish C, Mcdermott D. Relevance of genetics and genomics for prevention and treatment of cardiovascular disease: A scientific statement from the american heart association council on epidemiology and prevention, the stroke council, and the functional genomics and translational biology interdisciplinary working group. Circulation. 2007; 115:2878–2901. [PubMed: 17515457]

Borgan O, Goldstein L, Langholz B. Methods for the analysis of sampled cohort data in the Cox proportional hazards model. Ann. Statist. 1995:1749–1778.

Borgan O, Langholz B. Nonparametric estimation of relative mortality from nested case-control studies. Biometrics. 1993:593–602. [PubMed: 8369390]

Breslow N, Wellner J. Weighted Likelihood for Semiparametric Models and Two-phase Stratified Samples, with Application to Cox Regression. Scandinavian Journal of Statistics. 2007; 34:86–102.

Cai T, Zheng Y. Non-parametric Evaluation of Biomarker Accuracy under Nested Case-control Studie. J. Am. Stat. Assoc. 2011; 106:569–80. [PubMed: 22844169]

Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers in nested case-control studies. Biostatistics. 2012; 13:89–100. [PubMed: 21856652]

Chen K. Generalized case-cohort sampling. J. R. Statist. Soc. B. 2001; 63:791–809.

Chlebowski R, Johnson K, Kooperberg C, Pettinger M, Wactawski-Wende J, Rohan T, Rossouw J, Lane D, O'Sullivan M, Yasmeen S, et al. Calcium Plus Vitamin D Supplementation and the Risk of Breast Cancer. J Natl Cancer Inst. 2007; 100:1581–1591. [PubMed: 19001601]

Cox D. Regression models and life-tables. J. R. Statist. Soc. B. 1972:187–220.

D'Agostino R, Vasan R, Pencina M, Wolf P, Cobain M, Massaro J, Kannel W. General Cardiovascular Risk Profile for Use in Primary Care: The Framingham Heart Study. Circulation. 2008; 117:743–753. [PubMed: 18212285]

Davison, A.; Hinkley, D.; Canty, A. Bootstrap methods and their application. Cambridge University Press; 1999.

Garcia-Soidan PH, Hall P. On sample reuse methods for spatial data. Biometrics. 1997:273–281.

Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. Ann. Statist. 1992; 20:1903–28.

Gray R. Weighted analyses for cohort sampling designs. Lifetime data analysis. 2009; 15:24–40. [PubMed: 18712477]

Hall P, Horowitz JL, Jing B-Y. On blocking rules for the bootstrap with dependent data. Biometrika. 1995; 82:561–574.

Hankinson S, Willett W, Colditz G, Hunter D, Michaud D, Deroo B, Rosner B, Speizer F, Pollak M. Circulating concentrations of insulin-like growth factor I and risk of breast cancer. The Lancet. 1998; 351:1393–1396.

Härdle W, Horowitz J, Kreiss J-P. Bootstrap methods for time series. International Statistical Review. 2003; 71:435–459.

Hu F, Meigs J, Li T, Rifai N, Manson J. Inflammatory markers and risk of developing type 2 diabetes in women. Diabetes. 2004; 53:693–700. [PubMed: 14988254]

Ishibe N, Hankinson S, Colditz G, Spiegelman D, Willett W, Speizer F, Kelsey K, Hunter D. Cigarette smoking, cytochrome P450 1A1 polymorphisms, and breast cancer risk in the Nurses' Health study. Cancer research. 1998; 58:667–671. [PubMed: 9485019]

Jin Z, Ying Z, Wei L. A simple resampling method by perturbing the minimand. Biometrika. 2001; 88:381.

Karlson E, Chibnik L, Tworoger S, Lee I, Buring J, Shadick N, Manson J, Costenbader K. Biomarkers of inflammation and development of rheumatoid arthritis in women from two prospective cohort studies. Arthritis & Rheumatism. 2009; 60:641–652. [PubMed: 19248103]

Kosorok, M. Introduction to empirical processes and semiparametric inference. Springer Verlag: 2008.

Langholz B, Borgan Y. Estimation of absolute risk from nested case-control data. Biometrics. 1997; 53:767–74. [PubMed: 9192463]

Larson M, Atwood L, Benjamin E, Cupples L, D'Agostino R, Fox C, Govindaraju D, Guo C, Heard-Costa N, Hwang S, et al. Framingham Heart Study 100K project: genome-wide associations for cardiovascular disease outcomes. BMC medical genetics. 2007; 8:S5. [PubMed: 17903304]

Lloyd-Jones D, Nam B, D'Agostino R, Levy D, Murabito J, Wang T, Wilson P, O'Donnell C. Parental cardiovascular disease as a risk factor for cardiovascular disease in middle-aged adults a prospective study of parents and offspring. JAMA. 2004; 291:2204–2211. [PubMed: 15138242]

Lu W, Liu M. On estimation of linear transformation models with nested case-control sampling. Lifetime data analysis. 2012:1–14.

Ma J, Stampfer M, Giovannucci E, Artigas C, Hunter D, Fuchs C, Willett W, Selhub J, Hennekens C, Rozen R. Methylenetetrahydrofolate reductase polymorphism, dietary interactions, and risk of colorectal cancer. Cancer Research. 1997; 57:1098. [PubMed: 9067278]

Nan B, Yu M, Kalbfleisch J. Censored linear regression for case-cohort studies. Biometrika. 2006; 93:747–62.

Newey, W.; McFadden, D. Large Sample Estimation and Hypothesis Testing. In: McFadden, D.; Engler, R., editors. Handbook of Econometrics. Vol. IV. Amsterdam; North Holland: 1994. p. 2113-245.

Nomura A, Lee J, Stemmermann G, Combs G. Serum selenium and subsequent risk of prostate cancer. Cancer Epidemiology, Biomarkers & Prevention. 2000; 9:883–887.

Park Y, Wei L. Estimating subject-specific survival functions under the accelerated failure time model. Biometrika. 2003; 90:717–723.

Paynter N, Chasman D, Pare G, Buring J, Cook N, Miletich J, Ridker PM, Paynter N, Chasman D, Pare G, Buring J, Cook N, Miletich J, Ridker PM, Paynter N, Chasman D, Pare G, Buring J, Cook N, Miletich J, Ridker P. Association Between a Literature-Based Genetic Risk Score and Cardiovascular Events in Women. JAMA. 2010; 303:631–637. [PubMed: 20159871]

Pollard, D. Empirical processes: theory and applications. Institute of Mathematical Statistics; 1990.

Pradhan A, Manson J, Rossouw J, Siscovick D, Mouton C, Rifai N, Wallace R, Jackson R, Pettinger M, Ridker P. Inflammatory biomarkers, hormone replacement therapy, and incident coronary heart disease. JAMA. 2002; 288:980–987. [PubMed: 12190368]

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

Prentice R, Breslow N. Retrospective studies and failure time models. Biometrika. 1978; 65:153–158.

Rao J, Wu C. Resampling inference with complex survey data. Journal of the American Statistical Association. 1988; 83:231–241.

Rao J, Wu C, Yue K. Some recent work on resampling methods for complex surveys. Survey methodology. 1992; 18:209–217.

Ridker P, Buring J, Rifai N, Cook N. Development and validation of improved algorithms for the assessment of global cardiovascular risk in women: the Reynolds risk score. JAMA. 2007; 297:611–9. [PubMed: 17299196]

Ridker P, Rifai N, Rose L, Buring J, Cook N. Comparison of C-reactive protein and low-density lipoprotein cholesterol levels in the prediction of first cardiovascular events. N. Eng. J. Med. 2002; 347:1557–65.

Rundle A, Vineis P, Ahsan H. Design options for molecular epidemiology research within cohort studies. Cancer Epidemiology Biomarkers & Prevention. 2005; 14:1899.

Saarela O, Kulathinal S, Arjas E, Laäärä E. Nested case-control data utilized for multiple outcomes: a likelihood approach and alternatives. Statistics in medicine. 2008; 27:5991–6008. [PubMed: 18792086]

Salim A, Hultman C, Spareén P, Reilly M. Combining data from 2 nested case-control studies of overlapping cohorts to improve efficiency. Biostatistics. 2009; 10:70–79. [PubMed: 18550564]

Samuelsen S. A psudolikelihood approach to analysis of nested case-control studies. Biometrika. 1997; 84:379–394.

Samuelsen S, ANestad H, Skrondal A. Stratified case-cohort analysis of general cohort sampling designs. Scandinavian journal of statistics. 2007; 34:103–19.

Scheike T, Juul A. Maximum likelihood estimation for Cox's regression model under nested case-control sampling. Biostatistics. 2004; 5:193–206. [PubMed: 15054025]

Shao J. Impact of the bootstrap on sample surveys. Statistical Science. 2003; 18:191–198.

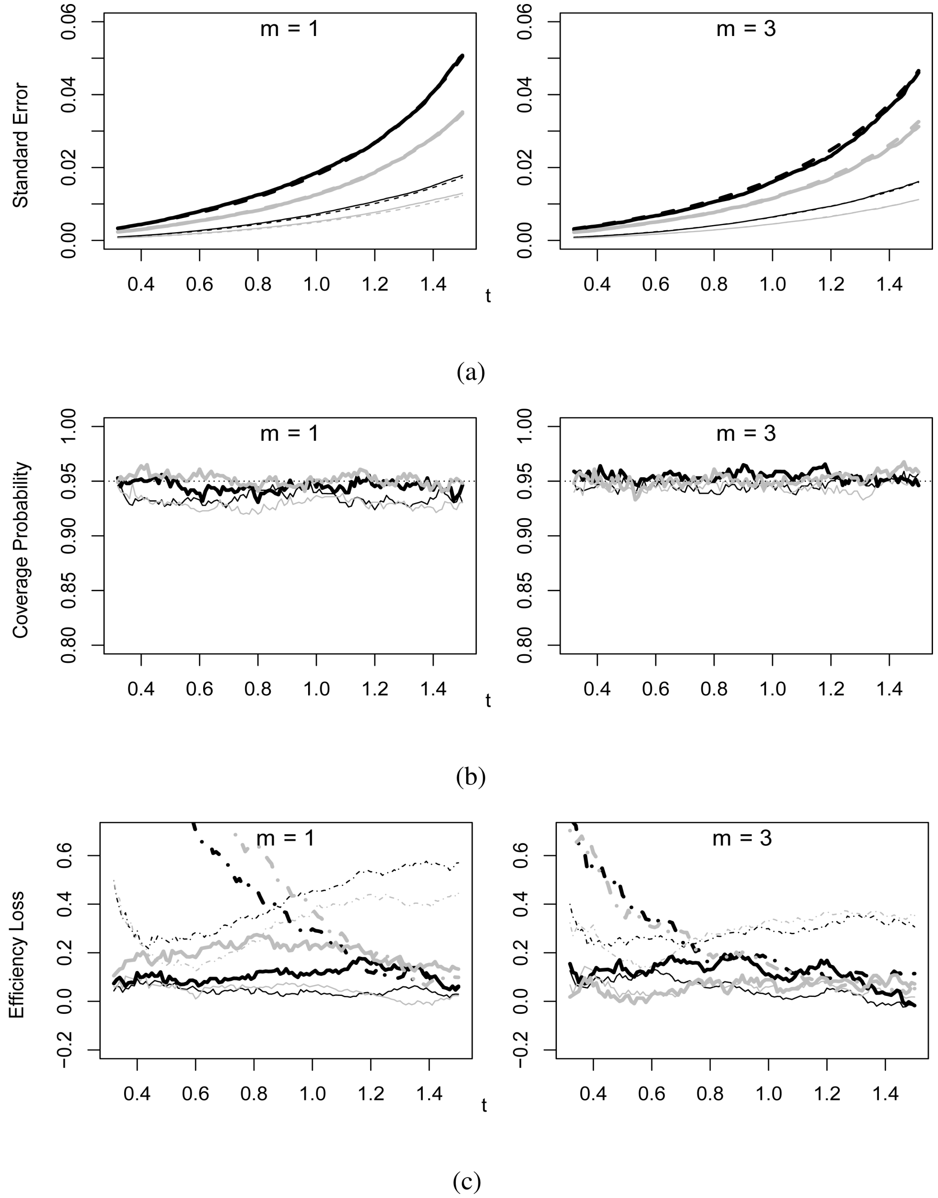
Sitter R. Comparing three bootstrap methods for survey data. Canadian Journal of Statistics. 1992; 20:135–154.

Thomas DC. Addendum to "Methods of cohort analysis: Appraisal by application to asbestos mining". Journal of the Royal Statistical Society, Series A, General. 1977; 140:483–485.

Wilson P, D'Agostino R, Levy D, Belanger A, Silbershatz H, Kannel W. Prediction of coronary heart disease using risk factor categories. Circulation. 1998; 97:1837–47. [PubMed: 9603539]

Wu K, Feskanich D, Fuchs C, Willett W, Hollis B, Giovannucci E. A Nested Case-Control Study of Plasma 25-Hydroxyvitamin D Concentrations and Risk of Colorectal Cancer. J Natl Cancer Inst. 2007; 99:1120–9. [PubMed: 17623801]

Zeng D, Lin D, Avery C, North K, Bray M. Efficient semiparametric estimation of haplotype-disease associations in case-cohort and nested case-control studies. Biostatistics. 2006; 7:486–502. [PubMed: 16500923]



(a)

(b)

(c)

**Figure 1.**

Simulation results for the survival function at marker level $W_0 = (0, 0)'$ (thinner curves) and $W_0 = (1, 1)'$ (thicker curves) under the Cox model without additional matching variables for sample sizes (i) n = 5000 (black curves) and (ii) n = 10000 (gray curves): (a) empirical (solid) and estimated (dashed) standard errors; (b) empirical coverage probabilities of the 95% confidence intervals; and (c) efficiency loss in the naive variance estimates (solid curves) and the conditional logistic regression based estimators (dot-dashed curves).

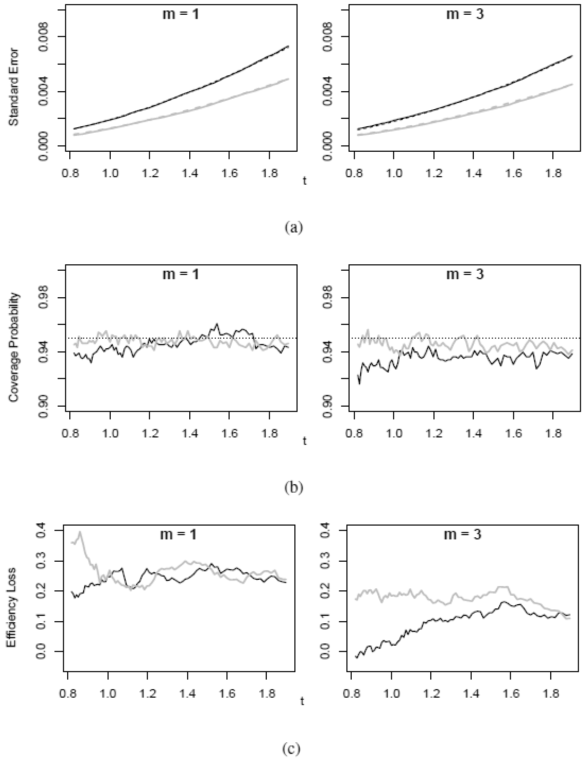


**Figure 2.**

Simulation results for the baseline survival function $\{S_{W_0}(t) : t \in [.8, 1.8]\}$ under the AFT model with two matching variables when sample sizes (i) n = 5000 (black curves) and (ii) n = 10000 (gray curves): (a) empirical (solid) and estimated (dashed) standard errors; (b) empirical coverage probabilities of the 95% confidence intervals; and (c) efficiency loss in the naive variance estimates.

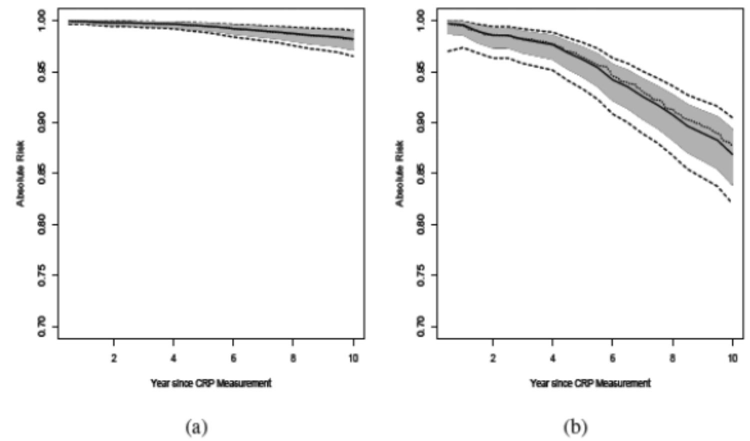


**Figure 3.**

Estimated absolute risks (solid lines) along with their 95% pointwise (shaded region) and simultaneous (dashed lines) CIs from the FOS based on the full cohort and an NCC design with m = 3 controls, and stratified by age and framingham score. (a) absolute risk for an individual with CRP and FR-score both at 10th percentiles respectively and gene score at 50th percentile of the population; (b) with absolute risk for an individual with CRP and FR-score both at 90th percentiles respectively and gene score at 50th percentile of the population. Shown also are the estimated absolute risks from the full cohort (dotted lines).

**Table 1**

Bias, empirical standard error (ESE), average of the estimated standard errors (ASE), and coverage probabilities (CovP) of the 95% CIs for $\beta = (.5, .5)'$ under the AFT model with two matching variables and dependent censoring.

| | | $\beta_1 = .5$ | | | | $\beta_2 = .5$ | | | |
|---|---|---|---|---|---|---|---|---|---|
| ***n*** | ***m*** | **Bias** | **ESE** | **ASE** | **CovP** | **Bias** | **ESE** | **ASE** | **CovP** |
| 5000 | 1 | −.002 | .086 | .082 | .933 | −.004 | .068 | .063 | .919 |
| 5000 | 3 | −.001 | .075 | .072 | .940 | −.003 | .057 | .056 | .943 |
| 10000 | 1 | .005 | .059 | .059 | .948 | −.003 | .046 | .045 | .940 |
| 10000 | 3 | .002 | .054 | .051 | .929 | −.002 | .039 | .039 | .951 |

## Table 2

Estimated coefficients (Est) and standard errors (SE) from multivariate Cox regression models from Framingham data based on the full cohort and an NCC design with m = 3 controls, and stratified by age and FR-score.

| | **Full cohort** | | **NCC** | |
|---|---|---|---|---|
| | **Est** | **SE** | **Est** | **SE** |
| Framingham score | .798 | .112 | .805 | .122 |
| log(CRP) | .500 | .085 | .513 | .104 |
| Gene score | −.153 | .235 | −.161 | .293 |