HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2021 August 15.

Published in final edited form as:

Stat Med. 2021 August 15; 40(18): 4035–4052. doi:10.1002/sim.9012.

Biomarker Evaluation Under Imperfect Nested Case–control Design

Xuan Wang[1], **Yingye Zheng**[2], **Majken Karoline Jensen**[3], **Zeling He**[1], **Tianxi Cai**[1,4]

[1]Department of Biostatistics, Harvard University, Boston, MA, USA

[2]Fred Hutchinson Cancer Research Center, Seattle, WA, USA

[3]Dept of Epidemiology, Harvard T.H. Chan School of Public Health, Boston, MA, USA

[4]Department of Biomedical Informatics, Harvard University, Boston, MA, USA

Summary

The nested case–control (NCC) design has been widely adopted as a cost-effective sampling design for biomarker research. Under the NCC design, markers are only measured for the NCC subcohort consisting of all cases and a fraction of the controls selected randomly from the matched risk sets of the cases. Robust methods for evaluating prediction performance of risk models have been derived under the inverse probability weighting (IPW) framework. The probabilities of samples being included in the NCC cohort can be calculated based on the study design[1] or estimated non-parametrically[2]. Neither strategy works well due to model mis-specification and the curse of dimensionality in practical settings where the sampling does not entirely follow the study design or depends on many factors. In this paper, we propose an alternative strategy to estimate the sampling probabilities based on a varying coefficient model, which attains a balance between robustness and the curse of dimensionality. The complex correlation structure induced by repeated finite risk set sampling makes the standard resampling procedure for variance estimation fail. We propose a perturbation resampling procedure that provides valid interval estimation for the proposed estimators. Simulation studies show that the proposed method performs well in finite samples. We apply the proposed method to the Nurses' Health Study II to develop and evaluate prediction models using clinical biomarkers for cardiovascular risk.

Keywords

finite population sampling; inverse probability weighting; nonparametric smoothing; resampling; risk prediction

1 | INTRODUCTION

Conducting rigorous biomarker validation studies is an important step in the translation of novel biomarkers into routine clinical practice for medical decision making. Such studies should follow design principles in sample selection to avoid bias[3]. Large prospective studies,

Correspondence Tianxi Cai. tcai@hsph.harvard.edu.

such as the Women's Health Initiative Study and the Nurses' Health Study, with exposures captured and biologic samples collected and stored prior to disease onset, can serve as a platform for biomarker research[4,5]. However, measuring biomarkers for large prospective cohorts is highly resource consuming. To make efficient use of stored samples from a cohort, two-phase sampling designs, including nested case–control (NCC) and case–cohort (CCH) studies, are often adopted as resource-efficient sampling strategies, especially when the outcome of interest is rare[6,7,8].

Under the NCC design, new markers are measured for all cases and a subset of controls randomly sampled without replacement from the risk sets of the cases. Sometimes controls are also matched to cases on variables such as gender and age. Many well-known biomarker studies nested in large cohorts have employed the NCC design[9,10,11, e.g.]. For example, in the Nurses' Health Study II (NHS_{II}), novel biomarkers, apoA1 concentration in whole plasma (WPA1) and concentration of apoE in whole plasma (apoE), were investigated for predicting the risk of Myocardial Infraction (MI)[11]. Due to limited resources and low incidence of MI, the biomarkers were measured on a nested case–control set, which included all cases and controls sampled from the 1:1 matched risk set of the cases with matching variables including smoking status, fasting status, age, and timing of blood collection.

To analyze NCC data, conditional logistic regression (CLR) model has traditionally been used when the focus is on the estimation of hazard ratio (HR) parameters, and sometimes the estimation of absolute risk parameters[12,13]. The CLR provides HR estimators under the Cox model from the full cohort, however cannot be extended to other models. Nor can the methods be used for estimating other parameters, such as the prediction accuracy parameters, which involve the distribution of the markers in the full cohort. For model parameter estimation, fully efficient maximum likelihood estimators (MLE) have also been proposed[14,15]. The MLE relies on the correct specification of the failure time model and requires that censoring is independent of the novel markers as well as additional modeling assumptions when there are multiple novel biomarkers and routine clinical variables measured on the full cohort.

As a flexible alternative, the inverse probability weighting (IPW) approach has been developed[16]. Recently, IPW estimators have also been developed for fitting models beyond the Cox model as well as for prediction performance measures including the receiver operating characteristic (ROC) curve, positive predictive value (PPV) and negative predictive value (NPV)[17,18,19]. Most existing IPW estimators for NCC studies calculates the true IPW (TIPW) sampling weights according to the study design and are consistent provided that the sampling weights are correctly obtained. However, the TIPW estimators may be invalid if the sampling is not implemented exactly according to the design. Such a scenario arises, for example, when the matching criteria are implemented only coarsely during the implementation of the sampling scheme due to practical concerns. Such an imperfect NCC design poses additional analytical challenges for estimating and evaluating risk prediction models. To overcome the bias, one may estimate the sampling weights non-parametrically via kernel smoothing as in Zheng et al.[2]. Obtaining such a non-parametric augmented IPW (NP) estimator, however, is not feasible when the number of matching variables is not very small due to the curse of dimensionality.

In this paper, we propose a new semi-non-parametric AIPW estimator, where the selection probability is estimated based on a flexible varying coefficient model. The AIPW estimator can incorporate a larger number of matching variables while remaining robust to the deviation from the intended sampling scheme. We derive the the asymptotic properties for the proposed estimators and come up with a resampling method to assess the variability of our proposed AIPW estimators.

The remainder of the paper is organized as follows. In Section 2, we provide model specification and describe the proposed point estimation procedures. Our proposed interval estimation procedure is given in Section 3. In Section 4, we report results of simulation studies to assess the finite sample performance of the proposed method. In Section 5, the data from NHS II is analyzed as illustration. Concluding remarks are given in Section 6. Theoretical studies of the proposed estimators are provided in the Appendix.

2 | ESTIMATING SAMPLING WEIGHTS VIA AIPW

Let T denote the survival outcome of interest and $\mathbf{Y} = (\mathbf{Y}_{\text{old}}^{\mathsf{T}}, \mathbf{Y}_{\text{new}}^{\mathsf{T}})^{\mathsf{T}}$ denote the vector of predictors for T, where $\mathbf{Y}_{\text{old}}$ denotes the vector of routine markers and $\mathbf{Y}_{\text{new}}$ denotes the vector of novel biomarkers. Due to censoring, T is only observed up to a bivariate vector $X = T \wedge C$ and $\delta = I(T \leq C)$, where C is the censoring time. Under the NCC design, $\mathbf{Y}_{\text{new}}$ is only measured if $V = 1$, where $V = \delta + (1 - \delta)V_0$ and V_0 is a binary indicator for whether a subject is sampled into the NCC subcohort as a control. We assume that the sampling of the controls is performed by matching to the cases according to a vector of matching variables $\mathbf{M}$. Suppose that the underling data for the full cohort consists of N independent and identically distributed random vectors, $\mathcal{D} = \{\mathbf{D}_i = (X_i, \delta_i, \mathbf{Y}_i^{\mathsf{T}}, \mathbf{M}_i^{\mathsf{T}})^{\mathsf{T}}, i = 1, \ldots, N\}$, while the observed data consist of $\mathcal{O} = \{\mathbf{O}_i = (X_i, \delta_i, \mathbf{Y}_{\text{old},i}^{\mathsf{T}}, V_i\mathbf{Y}_{\text{new},i}^{\mathsf{T}}, \mathbf{M}_i^{\mathsf{T}})^{\mathsf{T}}, i = 1, \ldots, N\}$. Let $\Omega = \{i : 1 \leq i \leq N\}$ and $\Omega_{\text{ncc}} = \{i : 1 \leq i \leq N,\ V_i = 1\}$ respectively denote the index sets for the full cohort and NCC subcohort.

Under the matched NCC design, for a case with event time X_i and matching variables $\mathbf{M}_i$, m controls are sampled from the matched risk set

$$\mathcal{R}_{\mathbf{W}_i} = \{k : X_k \geq X_i,\ |\ \mathbf{M}_k - \mathbf{M}_i\ | \leq \mathbf{a}_0\},$$

where $\mathbf{a}_0$ is a predetermined range vector and $\mathbf{W}_i = (\delta_i, X_i, \mathbf{M}_i^{\mathsf{T}}, \mathbf{Y}_{\text{old},i})^{\mathsf{T}}$. Let $\bar{\pi}_i = P(V_i = 1 \mid \mathcal{O})$ and $\bar{\pi}_{0i} = P(V_i = 1 \mid \mathcal{O}, \delta_i = 0)$ denote the true sampling probabilities under possibly imperfect NCC sampling. If the NCC sampling were implemented exactly accordingly to design, then $\bar{\pi}_i = \delta_i + (1 - \delta_i)\bar{\pi}_{0i}$ can be calculated as $\tilde{\pi}_i = \delta_i + (1 - \delta_i)\tilde{\pi}_{0i}$, where $\tilde{\pi}_{0i} = \tilde{\pi}_0(\mathbf{W}_i)$,

$$\tilde{\pi}_0(\mathbf{W}_i) = 1 - \prod_{j:\, j \in \mathcal{R}_{\mathbf{W}_i}} \left\{1 - \frac{m\delta_j}{|\ \mathcal{R}_{\mathbf{W}_j}\ | - 1}\right\}$$

and $|\mathscr{R}_{\mathbf{W}_i}|$ is the size of $\mathscr{R}_{\mathbf{W}_i}$[16]. Under the perfect NCC design, TIPW estimators can be constructed by weighting observations with the true weights $\tilde{\omega}_i = V_i / \tilde{\pi}_i$. To improve efficiency and robustness over the TIPW estimators, Zheng et al.[2] proposed NP estimators using non-parametrically estimated weights $\hat{\omega}_i^{\mathrm{NP}} = V_i / \hat{\pi}^{\mathrm{NP}}(\mathbf{W}_i)$, where $\hat{\pi}^{\mathrm{NP}}(\mathbf{W}_i) = \delta_i + (1-\delta_i)\hat{\pi}_0^{\mathrm{NP}}(\mathbf{W}_i)$,

$$\hat{\pi}_0^{\mathrm{NP}}(\mathbf{w}) = \frac{\sum_{i=1}^{N}(1-\delta_i)V_i K_b(\mathbf{W}_i - \mathbf{w})}{\sum_{i=1}^{N}(1-\delta_i)K_b(\mathbf{W}_i - \mathbf{w})}$$

is a non-parametric estimate of $\pi_0(\mathbf{w}) = P(V_i = 1 \mid \mathbf{W}_i = \mathbf{w}, \delta_i = 0)$, $K_b(\mathbf{w}) = b^{-q}\prod_{j=1}^{q} K(w_j / b)$, $K(\cdot)$ is a symmetric density function, and $b > 0$ denotes the bandwidth.

While the NP method can be used to incorporate imperfect NCC designs, it is infeasible when the dimension of $\mathbf{W}$ is not small. To overcome the limitations of TIPW and NP methods, we propose a semi-non-parametric AIPW method by approximating $\tilde{\pi}_{0i}$ via a flexible varying coefficient model

$$\pi_{0i} = g\{\boldsymbol{\beta}(\tilde{\pi}_{0i}, X_i)^{\top}\mathbf{Z}_i\} \text{ with } g(x) = \frac{e^x}{1+e^x} \tag{2.1}$$

where $\mathbf{Z}_i = (1, \boldsymbol{\Phi}_1(\mathbf{Y}_{\mathrm{old},i})^{\top}, \boldsymbol{\Phi}_2(\mathbf{M}_i)^{\top})^{\top}$, $\boldsymbol{\Phi}_1(\cdot)$ and $\boldsymbol{\Phi}_2(\cdot)$ are basis functions that allow potential non-linear effects, and $\boldsymbol{\beta}(\pi, x)$ is the unknown coefficient function. In practice, we find that the commonly used b-spline or natural splines basis with degree of freedom 3 works well. Equally spaced knots that covers most of the domain of the data are also desirable. We find that our results are not overly sensitive to the choice of the basis functions provided that they are reasonably flexible to capture non-linear effects. Under perfect NCC sampling, $\boldsymbol{\beta}(\tilde{\pi}_{0i}, X_i) = (g^{-1}(\tilde{\pi}_{0i}), \mathbf{0}^{\top})^{\top}$. On the other hand, when the sampling is imperfect, the flexible model provides accurate approximation to the true sampling probabilities while overcoming the curse of dimensionality associated with NP procedures.

To estimate $\boldsymbol{\beta}(\pi, x)$, we maximize a local logistic log-likelihood using observed data on $\{(V_i, \mathbf{Z}_i, X_i, \tilde{\pi}_{0i}) : \delta_i = 0\}$. Specifically, for any given (π, x), we estimate $\boldsymbol{\beta}(\pi, x)$ as $\hat{\boldsymbol{\beta}}(\pi, x)$, the solution to the estimating equation

$$\hat{\mathbf{U}}_{\pi,x}(\boldsymbol{\beta}) = N^{-1}\sum_{i=1}^{n} K_{\mathbf{b}}(\tilde{\pi}_{0i} - \pi, X_i - x)(1-\delta_i)\mathbf{Z}_i\left\{V_i - g(\boldsymbol{\beta}^{\top}\mathbf{Z}_i)\right\}$$

where $K_{\mathbf{b}}(\cdot) = (b_1 b_2)^{-1}K(\pi/b_1)K(x/b_2)$, $K(\cdot)$ is a symmetric density function function, $\mathbf{b} = (b_1, b_2)^{\top}$ is the bandwidth parameters vector which tend to 0 as $N \to \infty$. With $\hat{\boldsymbol{\beta}}(\pi, x)$, we estimate the sampling probability for the ith subject as

$$\hat{\pi}_i = \delta_i + (1-\delta_i)\hat{\pi}_{0i}, \text{ where } \hat{\pi}_{0i} = g\{\hat{\boldsymbol{\beta}}(\tilde{\pi}_{0i}, X_i)^\top \mathbf{Z}_i\}. \quad (2.2)$$

Then we construct our AIPW estimator using the augmented weights $\hat{\omega}_i = V_i / \hat{\pi}_i$. Under the correct specification of (2.1), we expect that $\max_{1 \le i \le N} | \hat{\pi}_i - \bar{\pi}_i | \to 0$ as $N \to \infty$.

3 | APPLICATION OF AIPW TO ROBUST RISK PREDICTION

In this section, we illustrate the application of the AIPW approach to developing and evaluating risk prediction models. Since one of the major goals of biomarker studies is to evaluate the predictive capacity of novel biomarkers, we consider quantifying the incremental value of $\mathbf{Y}_{\text{new}}$ in predicting T above and beyond routine markers $\mathbf{Y}_{\text{old}}$.

3.1 | Calibrated Risk Estimate

To predict risks based on $\mathbf{Y} = (\mathbf{Y}_{\text{old}}^\top, \mathbf{Y}_{\text{new}}^\top)^\top$ and $\mathbf{Y}_{\text{old}}$, we fit two proportional hazards (PH) models,

$$P(T \ge t \mid \mathbf{Y}) = S_{\text{all}}(t)^{\exp(\boldsymbol{\gamma}_{\text{all}}^\top \mathbf{Y})}, \quad (2.3)$$

$$P(T \ge t \mid \mathbf{Y}_{\text{old}}) = S_{\text{old}}(t)^{\exp(\boldsymbol{\gamma}_{\text{old}}^\top \mathbf{Y}_{\text{old}})}, \quad (2.4)$$

where $S_{\text{all}}(\cdot)$ and $S_{\text{old}}(\cdot)$ are unknown baseline survival functions and $\boldsymbol{\gamma}_{\text{all}}$ and $\boldsymbol{\gamma}_{\text{old}}$ are the corresponding log hazard ratio parameters. To estimate $\boldsymbol{\gamma}_{\text{all}}$ and $\boldsymbol{\gamma}_{\text{old}}$, we note that $\mathbf{Y}_{\text{new}}$ is only available for those in the NCC subcohort while $\mathbf{Y}_{\text{old}}$ is observed for all subjects. Thus, we propose to estimate $\boldsymbol{\gamma}_{\text{all}}$ by maximizing weighted log partial likelihood with AIPW weights $\hat{\omega}_i$:

$$\hat{\boldsymbol{\gamma}}_{\text{all}} = \text{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^{N} \hat{\omega}_i \delta_i \left[\boldsymbol{\gamma}^\top \mathbf{Y}_i - \log \left\{ \sum_{j=1}^{N} \hat{\omega}_j I(X_j \ge X_i) \exp(\boldsymbol{\gamma}^\top \mathbf{Y}_j) \right\} \right].$$

On the other hand, $\boldsymbol{\gamma}_{\text{old}}$ can be estimated as the standard maximum partial likelihood estimator, denoted by $\hat{\boldsymbol{\gamma}}_{\text{old}}$. It follows from Lin and Wei[20] and the consistency of the sampling probabilities that $\hat{\boldsymbol{\gamma}}_{\text{all}}$ and $\hat{\boldsymbol{\gamma}}_{\text{old}}$ respectively converge to deterministic vectors $\bar{\boldsymbol{\gamma}}_{\text{all}}$ and $\bar{\boldsymbol{\gamma}}_{\text{old}}$ as $N \to \infty$, regardless of the adequacy of the survival models (2.3) and (2.4). When models (2.3) and (2.4) hold, then $\bar{\boldsymbol{\gamma}}_{\text{all}} = \boldsymbol{\gamma}_{\text{all}}$ and $\bar{\boldsymbol{\gamma}}_{\text{old}} = \boldsymbol{\gamma}_{\text{old}}$.

To make a prediction for t-year survival, one may obtain a model-based estimate for $P(T \le t \mid \mathbf{Y})$ and $P(T \le t \mid \mathbf{Y}_{\text{old}})$ under (2.3) and (2.4). However, such a risk estimate may not be accurate under model mis-specifications. Following the calibrated risk prediction strategies proposed in Cai et al.[21], we predict t-year survival risk given $\mathbf{Y}$ and $\mathbf{Y}_{\text{old}}$ based on

$$\mathcal{S}_{\text{all}}(t \mid \boldsymbol{R}_{\text{all}}) \equiv \boldsymbol{P}(\boldsymbol{T} > t \mid \boldsymbol{R}_{\text{all}}) \quad \text{and} \quad \mathcal{S}_{\text{old}}(t \mid \boldsymbol{R}_{\text{old}}) \equiv \boldsymbol{P}(\boldsymbol{T} > t \mid \boldsymbol{R}_{\text{old}}),$$

respectively, where $\boldsymbol{R}_{\text{all}} = \mathbf{Y}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}_{\text{all}}$ and $\boldsymbol{R}_{\text{old}} = \mathbf{Y}_{\text{old}}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}_{\text{old}}$ are the limiting risk scores. The calibrated survival risk functions $\mathcal{S}_{\text{all}}(t \mid r)$ and $\mathcal{S}_{\text{old}}(t \mid r)$ can be non-parametrically estimated as $\widehat{\mathcal{S}}_{\text{all}}(t \mid r) = \exp\{-\widehat{\Lambda}_{\text{all}}(t \mid r)\}$ and $\widehat{\mathcal{S}}_{\text{old}}(t \mid r) = \exp\{-\widehat{\Lambda}_{\text{old}}(t \mid r)\}$, where

$$\widehat{\Lambda}_{\text{all}}(t \mid r) = \int_0^t \frac{\sum_i \widehat{\omega}_i K_h(\widehat{\boldsymbol{R}}_{\text{all},\,i} - r)\mathrm{d}N_i(u)}{\sum_i \widehat{\omega}_i K_h(\widehat{\boldsymbol{R}}_{\text{all},\,i} - r) I(X_i \geq u)}, \; \widehat{\Lambda}_{\text{old}}(t \mid r) = \int_0^t \frac{\sum_i K_h(\widehat{\boldsymbol{R}}_{\text{old},\,i} - r)\mathrm{d}N_i(u)}{\sum_i K_h(\widehat{\boldsymbol{R}}_{\text{old},\,i} - r) I(X_i \geq u)},$$

$\widehat{\boldsymbol{R}}_{\text{all},\,i} = \widehat{\boldsymbol{\gamma}}_{\text{all}}^{\mathsf{T}}\mathbf{Y}_i$, $\widehat{\boldsymbol{R}}_{\text{old},\,i} = \widehat{\boldsymbol{\gamma}}_{\text{old}}^{\mathsf{T}}\mathbf{Y}_{\text{old},\,i}$ and $N_i(t) = I(X_i \leq t)\delta_i$. The above calibrated risk prediction procedure essentially fits risk models (2.3) and (2.4) to summarize multi-variate risk markers into univariate risk scores, R_{all} and R_{old}, and then non-parametrically estimates the t-year risk given the risk score.

3.2 | Evaluating Prediction Performance

The accuracy of the risk prediction based on a given risk score R can be summarized by commonly used time dependent accuracy measures including the true positive rate (TPR), false positive rate (FPR), the receiver operating characteristic (ROC) curve, the positive predictive value (PPV), and the negative predictive value (NPV). These prediction performance measures typically consider the accuracy of a binary classification rule $R \geq r$ in predicting the t-year survival status $D_t = I(T \geq t)$. Specifically, the TPR and FPR of $R \geq r$ in prediction D_t are respectively defined as

$$\text{TPR}(r \mid t) = \boldsymbol{P}(\boldsymbol{R} \geq r \mid \boldsymbol{T} < t), \text{ and } \text{FPR}(r \mid t) = \boldsymbol{P}(\boldsymbol{R} \geq r \mid \boldsymbol{T} \geq t).$$

The ROC curve, $\text{ROC}(u|t) = \text{TPR}\{\text{FPR}^{-1}(u|t)|t\}$, summarizes the trade-off between the FPR and TPR as the cut-off value varies. The PPV and NPV of $R \geq r$ are defined as

$$\text{PPV}(t \mid r) = \boldsymbol{P}(\boldsymbol{T} < t \mid \boldsymbol{R} \geq r), \text{ and } \text{NPV}(t \mid r) = \boldsymbol{P}(\boldsymbol{T} \geq t \mid \boldsymbol{R} < r).$$

To estimate the prediction accuracy for R_{all} and R_{old}, we note that all the aforementioned parameters are functionals of $\mathcal{S}_{\text{all}}(t \mid r)$, $\mathcal{S}_{\text{old}}(t \mid r)$, $\mathcal{F}_{\text{all}}(r) = P(\boldsymbol{R}_{\text{all}} \leq r)$ and $\mathcal{F}_{\text{old}}(r) = P(\boldsymbol{R}_{\text{old}} \leq r)$. For example, the TPR and FPR of $R_{\text{all}} \geq r$ can be respectively written as

$$\text{TPR}_{\text{all}}(r \mid t) = \frac{1 - \mathcal{F}_{\text{all}}(r) - \int_r^{\infty} \mathcal{S}_{\text{all}}(t \mid u) d\mathcal{F}_{\text{all}}(u)}{1 - \int \mathcal{S}_{\text{all}}(t \mid u) d\mathcal{F}_{\text{all}}(u)}, \text{ and } \text{FPR}_{\text{all}}(r \mid t) = \frac{\int_r^{\infty} \mathcal{S}_{\text{all}}(t \mid u) d\mathcal{F}_{\text{all}}(u)}{\int \mathcal{S}_{\text{all}}(t \mid u) d\mathcal{F}_{\text{all}}(u)}.$$

The trade-off between $\mathrm{TPR}_{\mathrm{all}}(\cdot \mid t)$ and $\mathrm{FPR}_{\mathrm{all}}(\cdot \mid t)$ can be summarized based on the receiver operating characteristic (ROC) curve $\mathrm{ROC}_{\mathrm{all}}(u \mid t) = \mathrm{TPR}_{\mathrm{all}}\{\mathrm{FPR}_{\mathrm{all}}^{-1}(u \mid t) \mid t)\}$, where u is any specified FPR level of interest. The marginal distribution functions $\mathscr{F}_{\mathrm{all}}(r)$ and $\mathscr{F}_{\mathrm{old}}(r)$ can be respectively estimated as

$$\widehat{\mathscr{F}}_{\mathrm{all}}(r) = \frac{\sum_{i=1}^{N} \widehat{\omega}_i I(\widehat{\boldsymbol{R}}_{\mathrm{all},\, i} \geq r)}{\sum_{i=1}^{N} \widehat{\omega}_i}, \text{ and } \widehat{\mathscr{F}}_{\mathrm{old}}(r) = N^{-1} \sum_{i=1}^{N} \boldsymbol{I}(\widehat{\boldsymbol{R}}_{\mathrm{old},\, i} \geq r).$$

Subsequently, we may construct plug-in estimators for $\mathrm{TPR}_{\mathrm{all}}(\cdot \mid t)$ and $\mathrm{FPR}_{\mathrm{all}}(\cdot \mid t)$ as

$$\widehat{\mathrm{TPR}}_{\mathrm{all}}(r \mid t) = \frac{1 - \widehat{\mathscr{F}}_{\mathrm{all}}(r) - \int_r^{\infty} \widehat{\mathscr{S}}_{\mathrm{all}}(t \mid u) d\widehat{\mathscr{F}}_{\mathrm{all}}(u)}{1 - \int \widehat{\mathscr{S}}_{\mathrm{all}}(t \mid u) d\widehat{\mathscr{F}}_{\mathrm{all}}(u)} \text{ and } \widehat{\mathrm{FPR}}_{\mathrm{all}}(r \mid t) = \frac{\int_r^{\infty} \widehat{\mathscr{S}}_{\mathrm{all}}(t \mid u) d\widehat{\mathscr{F}}_{\mathrm{all}}(u)}{\int \widehat{\mathscr{S}}_{\mathrm{all}}(t \mid u) d\widehat{\mathscr{F}}_{\mathrm{all}}(u)},$$

respectively. Similar plug-in estimators can be constructed for other accuracy parameters. We may quantify the incremental value (IncV) of $\mathbf{Y}_{\mathrm{new}}$ based on the difference between the accuracy of R_{all} and R_{old}. For example, the IncV of $\mathbf{Y}_{\mathrm{new}}$ with respect to the ROC curve at FPR level of u_0 can be estimated as $\widehat{\mathrm{ROC}}_{\mathrm{all}}(u_0 \mid t) - \widehat{\mathrm{ROC}}_{\mathrm{old}}(u_0 \mid t)$, where $\widehat{\mathrm{ROC}}_{\mathrm{all}}(u_0 \mid t) = \widehat{\mathrm{TPR}}_{\mathrm{all}}\{\widehat{\mathrm{FPR}}_{\mathrm{all}}^{-1}(u_0 \mid t) \mid t\}$ and $\widehat{\mathrm{ROC}}_{\mathrm{old}}$ is the estimated ROC curve for R_{old}.

3.3 | Resampling Based Interval Estimation

To estimate the asymptotic variance of the proposed AIPW estimators, we propose a perturbation resampling procedure. Specifically, let $\mathbf{I} = (I_1, \ldots, I_N)^{\top}$ be a vector of independent and identically distributed non-negative random variables with mean 1 and variance 1. We first obtain perturbed counterpart of $\widehat{\boldsymbol{\beta}}(\pi, x)$ as $\widehat{\boldsymbol{\beta}}^*(\pi, x)$, the solution to the estimating equation

$$\widehat{\mathbf{U}}^*_{\pi,\, x}(\boldsymbol{\beta}) = N^{-1} \sum_{i=1}^{n} \boldsymbol{K}_{\mathbf{b}}(\widetilde{\pi}_{0i} - \pi, X_i - x)(1 - \delta_i)\mathbf{Z}_i \left\{ V_i - g(\boldsymbol{\beta}^{\top}\mathbf{Z}_i) \right\} \boldsymbol{I}_i.$$

Then we perturb the AIPW weights as

$$\widehat{\omega}_i^* = \left\{ \delta_i + (1 - \delta_i) \frac{V_{0i}}{\widehat{\pi}_{0i}^*} \right\} I_i \text{ with } \widehat{\pi}_{0i}^* = g\{(\widehat{\boldsymbol{\beta}}^*(\widetilde{\pi}_{0i}, X_i)^{\top}\mathbf{Z}_i\}.$$

Subsequently, we perturb all AIPW estimators by replacing $\widehat{\omega}_i$ with $\widehat{\omega}_i^*$. Specifically, we perturb $\widehat{\boldsymbol{\gamma}}_{\mathrm{all}}$ as

$$\widehat{\boldsymbol{\gamma}}^*_{\mathrm{all}} = \mathrm{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^{N} \widehat{\omega}_i^* \delta_i \left[\boldsymbol{\gamma}^{\top}\mathbf{Y}_i - \log \left\{ \sum_{j=1}^{N} \widehat{\omega}_j^* I(X_j \geq X_i) \exp(\boldsymbol{\gamma}^{\top}\mathbf{Y}_j) \right\} \right],$$

and perturb $\widehat{\mathcal{S}}_{\text{all}}(t \mid r)$ as $\widehat{\mathcal{S}}^*_{\text{all}}(t \mid r) = \exp\{-\widehat{\Lambda}^*_{\text{all}}(t \mid r)\}$, where

$$\widehat{\Lambda}^*_{\text{all}}(t \mid r) = \int_0^t \frac{\sum_i \widehat{\omega}_i^* K_h(\widehat{\boldsymbol{R}}^*_{\text{all},i} - r)\mathrm{d}N_i(u)}{\sum_i \widehat{\omega}_i^* K_h(\widehat{\boldsymbol{R}}^*_{\text{all},i} - r) I(X_i \geq u)}, \text{ and } \widehat{\boldsymbol{R}}^*_{\text{all},i} = \mathbf{Y}^{\mathsf{T}}_{\text{all},i}\widehat{\boldsymbol{\gamma}}^*_{\text{all}}.$$

The accuracy parameters can be perturbed similarly. For example, we may obtain

$$\widehat{\text{TPR}}^*_{\text{all}}(r \mid t) = \frac{1 - \widehat{\mathcal{F}}^*_{\text{all}}(r) - \int_r^\infty \widehat{\mathcal{S}}^*_{\text{all}}(t \mid u) d\widehat{\mathcal{F}}^*_{\text{all}}(u)}{1 - \int \widehat{\mathcal{S}}^*_{\text{all}}(t \mid u) d\widehat{\mathcal{F}}^*_{\text{all}}(u)},$$

where $\widehat{\mathcal{F}}^*_{\text{all}}(r) = \sum_{i=1}^N \widehat{\omega}_i^* I(\widehat{\boldsymbol{R}}^*_{\text{all},i} \leq r) / (\sum_{i=1}^N \widehat{\omega}_i^*)$.

For IncV parameters, the estimation of model parameters related to the reduced model only involve full cohort data and thus the perturbation will only involve weighting observations by $\{I_i\}$. Specifically, $\widehat{\boldsymbol{\gamma}}_{\text{old}}$ is perturbed as

$$\widehat{\boldsymbol{\gamma}}^*_{\text{old}} = \text{argmax}_{\boldsymbol{\gamma}} \sum_{i=1}^N I_i \delta_i \left[\boldsymbol{\gamma}^{\mathsf{T}} \mathbf{Y}_{\text{old},i} - \log\left\{\sum_{j=1}^N I_j I(X_j \geq X_i) \exp(\boldsymbol{\gamma}^{\mathsf{T}} \mathbf{Y}_{\text{old},j})\right\}\right],$$

and $\widehat{\mathcal{S}}^*_{\text{old}}(t \mid r) = \exp\{-\widehat{\Lambda}^*_{\text{old}}(t \mid r)\}$, where

$$\widehat{\Lambda}^*_{\text{old}}(t \mid r) = \int_0^t \frac{\sum_i I_i K_h(\widehat{\boldsymbol{R}}^*_{\text{old},i} - r)\mathrm{d}N_i(u)}{\sum_i I_i K_h(\widehat{\boldsymbol{R}}^*_{\text{old},i} - r) I(X_i \geq u)}, \text{ and } \widehat{\boldsymbol{R}}^*_{\text{old},i} = \mathbf{Y}^{\mathsf{T}}_{\text{old},i}\widehat{\boldsymbol{\gamma}}^*_{\text{old}}.$$

Similar strategies can be used for accuracy parameters such as $\widehat{\text{TPR}}^*_{\text{old}}(c \mid t)$ and $\widehat{\text{FPR}}^*_{\text{old}}(c \mid t)$.

To obtain variance estimators and construct confidence intervals, we may obtain a large number, say B, of realizations of $\mathbf{I}$. For each realization of $\mathbf{I}$, we obtain the above perturbed estimates. The empirical distribution of the B sets of perturbed estimates can be used for inference. For example, the empirical variance of $\widehat{\text{ROC}}^*_{\text{all}}(u_0 \mid t) - \widehat{\text{ROC}}^*_{\text{old}}(u_0 \mid t)$ can be used to approximate the variance of $\widehat{\text{ROC}}_{\text{all}}(u_0 \mid t) - \widehat{\text{ROC}}_{\text{old}}(u_0 \mid t)$.

4 | NUMERICAL STUDIES

We performed extensive simulations to evaluate the finite sample performance of the proposed estimators and to compare with other estimators under NCC design when the design is carried out perfectly or imperfectly. We generate $\mathbf{Y} = (Y_{\text{old}}, Y_{\text{new}})^{\mathsf{T}}$ from a bivariate normal distribution with zero mean, unit variance and correlation 0.5. Given $\mathbf{Y}$, we generate T from a PH model

$$P(T \geq t \mid \mathbf{Y}) = \exp\left[-\exp\left\{\log(0.01t) + \log(2)\boldsymbol{Y}_{\text{new}} + \log(3)\boldsymbol{Y}_{\text{old}}\right\}\right].$$

The censoring time was generated from two settings: (I) $C \sim C_{\text{IND}} = \min(C_a, C_b)$, where C_a ~ Uniform(0.5, 2) and $C_b \sim$ Gamma(shape = 2, rate = 2); (II) $\boldsymbol{C} \sim \boldsymbol{C}_{\text{DEP}} = \min\{\boldsymbol{C}_a, \boldsymbol{C}_b'(\mathbf{Y})\}$, where $\boldsymbol{C}_b'(\mathbf{Y}) = \exp\{-(\boldsymbol{Y}_{\text{new}} + \boldsymbol{Y}_{\text{old}})/5\} + 0.5$. This leads to covariate independent censoring in (I) and covariate dependent censoring in (II). The censoring rate and event rate (proportion of cases) are around 15% and 5%, respectively. We let $N = 2000$, and selected the NCC cohort by including all the cases and $m = 1$ control per case. Under each configuration, results were summarized based on 500 simulated datasets. We obtain estimators for $\bar{\boldsymbol{\gamma}}_{\text{all}} = (\bar{\gamma}_1, \bar{\gamma}_2)$ in model (2.3) and TPR_{u_0}, PPV_{u_0}, NPV_{u_0} at FPR= u_0, with u_0 taken to be 0.05, 0.1, 0.2. We also compared the proposed approach with existing methods including the TIPW estimator of Cai and Zheng[19], NP estimators of Zheng et al.[2] and conditional logistic regression method based estimator, denoted as CLR.

We considered three settings. In the first setting, setting (1), the matching covariate vector $\mathbf{M} = (M_1, M_2)$ with matching window $a_0 = (0, 0)$, where $M_1 = \sum_{l=1}^{2} I(\boldsymbol{Y}_{\text{old}} \leq y_{ql})$, y_q was the $100q$th percentiles of Y_{old} and $q_1 = 0.33$, $q_2 = 0.66$, $M_2 \sim$ Bernouli(0.5); In setting (2), matching variable $\mathbf{M} = (M_1, \ldots, M_5)^\top$ with matching window $a_0 = (0, 2, 2, 5, 0)$, where M_1 is the same as in setting (1), $M_2 \sim \lfloor 0.3e^{\mathcal{N}} \rfloor$, $M_3 \sim \lfloor 5\phi(\boldsymbol{Y}_{\text{old}} + \mathcal{N}) \rfloor$, $M_4 \sim \lfloor \text{Uniform}(0, 10) \rfloor$, and $M_5 \sim$ Bernouli(0.5), ϕ is a normal density function, and $\mathcal{N} \sim N(0, 1)$; In setting (3), matching variable $\mathbf{M} = (M_1, M_2, M_3, M_4)^\top$ with a varying window in that we intend to match with window $a_0 = (0, 0, 0, 0)$ but when the number of subjects is not sufficient in the risk set for some cases, we relax the criterion to matching window $a = (0, 0, 2, 2)$ to select controls in the new risk set. Here M_1 is the same as in setting (1), $M_2 = I(\boldsymbol{Y}_{\text{old}} + \mathcal{N} > 0)$ with $\mathcal{N} \sim N(0, 1)$, $M_3 \sim \lfloor 5\phi(\boldsymbol{Y}_{\text{old}} + \mathcal{N}) \rfloor$ and $M_4 \sim \lfloor 0.2e^{\mathcal{N}} \rfloor$.

Results summarizing the performance of the proposed point and interval estimators across settings (1) - (3) are presented in Table 1-3. The point estimators have negligible biases. The average of the standard errors (ASEs) are close to the corresponding empirical standard errors (SEs), and the empirical coverage probabilities (CP) of the 95% confidence intervals are close to the nominal level. These results confirm the validity of the proposed estimation procedures in finite sample.

In setting (1), sampling is correctly carried out and **M** is low dimensional, and hence all three methods (TIPW, AIPW, NP) are valid. As shown in Table 1, all three estimators have negligible biases, TIPW and NP have comparable efficiency with respect to mean squared error (MSE), and AIPW is a little more efficient than the TIPW and NP estimators. In setting (2), the sampling is carried out correctly but the matching variable is of a higher dimension, which leads to curse of dimensionality for the NP method while the TIPW remains valid. As shown in Table 2, the TIPW and AIPW both have negligible biases, while the NP exhibits higher biases. Setting (3) is a commonly encountered imperfect NCC sampling setting that is similar to the motivating example of the NHS II study. In this case, the TIPW estimator is

biased as expected. There is also bias observed for the NP estimators due to the curse of dimensionality, whereas the AIPW estimator still maintains negligible bias. In addition, the AIPW estimator is substantially more efficient than both the TIPW and NP estimators with respect to MSE, with relative efficiency as high as 6 compared to the TIPW estimator and 5 compared to the NP estimator. In all the settings considered, the CLR estimator is either more biased or less efficient compared to other estimators.

To examine whether our proposed method performs well under settings with a very low event rate, we also generated data under a slight variation of the above PH model with a substantially lower baseline hazard leading to about 0.5% of event rate and independent censoring. We sampled the NCC cohort under setting (3) and obtained estimates as above. As shown in Table 4, the proposed AIPW estimates have small biases and high relative efficiencies.

5 | REAL DATA ANALYSIS

High-density lipoprotein (HDL) is a protein-lipid complex which carries a range of proteins. These proteins differ in size and structure, which determines the functional properties and metabolism of HDL[22]. The plasma total apoA-1 concentration (WPA1) is well known to be strongly and consistently predictive of cardiovascular risk[23]. In addition to apoA-1, HDL also contains other proteins including apoA2, apoC3 and apoE. ApoC3, present on 8-15% of HDL particles, has been shown to be associated with the risk of obesity and diabetes[24,25]. To assess the predictiveness of these lipoprotein markers for the risk of developing myocardial infarction (MI), an NCC biomarker study was performed within the NHS II blood cohort consisting of 29,240 registered nurses enrolled around 1989[26]. Among participants who were free of diagnosed cardiovascular disease or cancer at blood draw, 144 women were identified in the cohort with incident MI between blood draw and January 2016. Using a risk-set sampling, 144 controls were to be selected randomly and 1:1 matched on age, fasting (yes, no), smoking (never, past, current <15 cigarettes/day, current > 15 cigarettes/day, resulting in three dummy variables), and month of blood drawn. However, due to the lack of samples satisfying the matching criteria and having sufficient stored plasma for biomarker quantification, NCC design was not followed exactly during the control sampling process, yielding an imperfect NCC design. If the matching criteria is followed, the matching window should be $a_0 = (0, 0, 0, 0, 2, 2)$. But if for some case, there is no control in its risk set, the matching criteria is relaxed but not known. For example, the age difference maybe relaxed to 5 years so that there are controls to select from for this case.

The outcome of interest is the time from blood drawn to diagnosis of MI. For an individual without an event, failure time was censored at the earlier date between the last contact date and January 2016. Routine risk factors included smoking, age, diabetes, high cholesterol, and medication for HBP. These factors are available from the full cohort. Measures of the new biomarkers, WPA1 and apoE, are only available for the NCC subcohort. To account for the subcohort sample, we fitted a weighted Cox PH model including WPA1 and apoE and other baseline clinical variables as covariates using the data from the NCC subset. Since the sampling depends on many levels of covariates, it was difficult to estimate the weights using the NP approach (Exiting packages for nonparametric estimation of the selection probability

all failed). Due to the additional adjustment in matching criteria, the 'true' weights were not retrievable. Therefore, we calculated the weights using the proposed AIPW techniques. As presented in Table 5, more frequent smoking (>15 cig/d), having diabetes or high cholesterol, or medication for high blood pressure, and high values of apoE are significantly associated with high risk of MI. In particular apoE predicts the time to MI beyond clinical factors, with an HR of 1.427 (95% CI: 1.140, 1.786). We also considered fitting the Cox model using the weights calculated strictly from the original protocol, the IPW method. For the variable more frequent smoking (>15 cig/d) versus never smoking, the estimated HR by the AIPW method is significantly above 1 while not significantly above 1 by the IPW method, which does not reflect the findings based on the existing literature. This is as expected, as the weights in this situation do not accurately account for the sampling procedures actually implemented, and this might potentially lead to biased estimates in the main regression model. The results highlight the importance of robust procedures in the calculation of the sampling weights, though the difference between the IPW and AIPW estimators is less pronounced for new markers. The estimated effects of CLR are a little different from the IPW and AIPW estimators.

We then calculated the in-sample accuracy measures of the model scores for predicting risk of MI by 158 months ($t = 158$) using the proposed method. The estimates of TPR, PPV, NPV at FPR=0.05, 0.1, 0.2 and AUC for the Cox model with baseline covariates as well as WPA1 and apoE are listed in Table 6 along with the IncV of the corresponding accuracy measures compared to the performance of a Cox model without the biomarkers. Results show that adding WPA1 and apoE to the Cox model with baseline covariates leads to no significant improvement in the accuracy measures, though apoE has a significant association with time to MI.

6 | DISCUSSION

Cost-effective two-phase sampling designs have been widely adopted in biomarker research in recent years. The nonrandom sampling of the NCC designs introduces complex data structures, which should be dealt with carefully to avoid bias. One well-recognized barrier in the analysis of two-phase designs is that the control selection procedures are often complicated in practical implementation: many matching factors are considered, and the window of selection for each variable might be adjusted in an ad-hoc fashion over the course of study, making it infeasible to retrieve the 'true' sampling weights. Robust nonparametric procedures for estimating the weights can consistently recover the weights according to the actual sampling, however they are limited in handling more than a few matching factors. In the case that the number of matching variables and routine markers Y_{old} exceed 5, the NP method of Zheng et al.[2] often becomes infeasible both theoretically and practically. On the other hand, our proposed AIPW method leverages the true sampling weights as a reasonable starting point and uses a sufficiently flexible model to estimate the effect on sampling of both variables involved in control selection and other correlated variables. Compared to the NP approach, the proposed AIPW procedure is able to incorporate a larger number of variables to augment the weights, while maintaining reasonable robustness and efficiency. It is important to note that matching on a large number of variables is generally not desirable

since it inherently increases the chance of the matched risk sets being empty. We therefore do not recommend that in practice.

There are a couple of future directions/limitations in this line of research. The approach we proposed can easily be extended to other types of two-phase sampling such as a covariate-stratified case–cohort studies. Flexible methods are also needed to account for other practical complications in two-phase sampling. Our methods here assume a NCC study where all cases will be selected due to a low incidence rate. However in practice, due to cases and sample availability, not all cases can be sampled[27]. This may complicate the inference procedure and warrants future research.

The R code for carrying out the proposed AIPW procedure is available upon request.

ACKNOWLEDGEMENT

This research were funded by U01CA86368 and R01CA236558 awarded by the National Institutes of Health.

APPENDIX

APPENDIX A.

Note that in the appendixes, the derivations are with respect to the whole data and the proposed AIPW estimator, so we omit the subscript ‘all’ for notation convenience.

In this section, we show the asymptotic normality of the proposed AIPW estimator.

Assume C has a finite support $[0, \tau]$, $P(T > \tau) > 0$ and the markers $\mathbf{Y}$ are continuous and bounded. The limit of $\hat{\boldsymbol{\gamma}}$, which is $\bar{\boldsymbol{\gamma}}$, is in the interior of a compact parameter space $\Omega_{\boldsymbol{\gamma}}$. Suppose the regularity conditions in Andersen and Gill[28] hold. Similarly to Du and Akritas[29], we assume the kernel function K is a symmetric probability density function with finite support and bounded second derivative. In addition, we assume the joint density of $\boldsymbol{R} = \mathbf{Y}^{\mathsf{T}}\bar{\boldsymbol{\gamma}}$, T, and C has continuous derivatives.

Denote $\beta_i = \beta(\tilde{\pi}_{0i}, X_i)$, we first get the asymptotic expression of $N^{1/2}(\hat{\beta}_i - \beta_i)$, which will be used in later derivations. Recalling that

$$\widehat{\mathbf{U}}_{\tilde{\pi}_{0i}, X_i}(\beta_i) = \frac{1}{N}\sum_{j=1}^{N}(1-\delta_j)[V_j - \exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j) / \{1+\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)\}]\mathbf{Z}_j K_b((\tilde{\pi}_{0i}, X_i) - (\tilde{\pi}_{0j}, X_j)).$$

The derivative of $\widehat{\mathbf{U}}_{\tilde{\pi}_{0i}, X_i}(\beta_i)$ with respect to $\boldsymbol{\beta}_i$ is

$$\frac{\partial \widehat{\mathbf{U}}_{\tilde{\pi}_{0i}, X_i}(\beta_i)}{\partial \beta_i} = -\frac{1}{N}\sum_{j=1}^{N}\frac{\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)}{\{1+\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)\}^2}\mathbf{Z}_j\mathbf{Z}_j^{\mathsf{T}}(1-\delta_j)K_b((\tilde{\pi}_{0i}, X_i) - (\tilde{\pi}_{0j}, X_j)),$$

which converges to $-\Sigma_i := -\tilde{\pi}_{0i}(1-\tilde{\pi}_{0i})E[\mathbf{Z}_j\mathbf{Z}_j^{\mathsf{T}} \mid \tilde{\pi}_{0j}=\tilde{\pi}_{0i}, X_j=X_i]f(\tilde{\pi}_{0i}, X_i)$, where $f(\cdot,\cdot)$ is the density function of $(\tilde{\pi}_{0i}, X_i)$. It follows that

$$N^{1/2}(\hat{\beta}_i-\beta_i)=\Sigma_i^{-1}N^{-1/2}\sum_{j=1}^{N}[V_j-\frac{\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)}{1+\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)}]\mathbf{Z}_j(1-\delta_j)K_b((\tilde{\pi}_{0i}, X_i) - (\tilde{\pi}_{0j}, X_j))+o_p(1)\,. \tag{A.1}$$

For the proposed AIPW estimators with a general form

$$\hat{U}=N^{-1/2}\sum_{i=1}^{N}\hat{\omega}_i U_i, \tag{A.2}$$

where $E(U_i)=0$, $\hat{\omega}_i=V_i/\hat{\pi}_i$ and $\hat{\pi}_i=\delta_i+(1-\delta_i)\hat{\pi}_{0i}$, we have

$$\hat{U}=N^{-1/2}\sum_{i=1}^{N}\hat{\omega}_i U_i=N^{-1/2}\sum_{i=1}^{N}U_i+N^{-1/2}\sum_{i=1}^{N}(\tilde{\omega}_i-1)U_i+N^{-1/2}\sum_{i=1}^{N}(\hat{\omega}_i-\tilde{\omega}_i)U_i \equiv I_1+I_2+I_3,$$

where $I_3=N^{-1/2}\sum_{i=1}^{N}V_i(\frac{1}{\hat{\pi}_{0i}}-\frac{1}{\tilde{\pi}_{0i}})U_i=-N^{-1/2}\sum_{i=1}^{N}V_i\frac{\hat{\pi}_{0i}-\tilde{\pi}_{0i}}{\hat{\pi}_{0i}\tilde{\pi}_{0i}}U_i$

$$=-N^{-1/2}\sum_{i=1}^{N}\tilde{\omega}_i\frac{\hat{\pi}_{0i}-\tilde{\pi}_{0i}}{\hat{\pi}_{0i}}U_i=-N^{-1/2}\sum_{i=1}^{N}\tilde{\omega}_i U_i\frac{\hat{\pi}_{0i}-\tilde{\pi}_{0i}}{\tilde{\pi}_{0i}}+o_p(1)$$

$$=-N^{-1/2}\sum_{i=1}^{N}\tilde{\omega}_i U_i\frac{\dot{g}(\beta_i^{\mathsf{T}}\mathbf{Z}_i)}{g(\beta_i^{\mathsf{T}}\mathbf{Z}_i)}\mathbf{Z}_i^{\mathsf{T}}(\hat{\beta}_i-\beta_i)+o_p(1)$$

$$=-N^{-1}\sum_{i=1}^{N}\tilde{\omega}_i U_i(1-\tilde{\pi}_{0i})\mathbf{Z}_i^{\mathsf{T}}\Sigma_i^{-1}N^{-1/2}\sum_{j=1}^{N}(V_j-\frac{\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)}{1+\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)})\mathbf{Z}_j(1-\delta_j)K_b((\tilde{\pi}_{0i}, X_i)-(\tilde{\pi}_{0j}, X_j))+o_p(1)$$

$$=-N^{-1/2}\sum_{j=1}^{N}E[U_i\mathbf{Z}_i^{\mathsf{T}} \mid \tilde{\pi}_{0i}=\tilde{\pi}_{0j}, X_i=X_j]E[\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}} \mid \tilde{\pi}_{0i}=\tilde{\pi}_{0j}, X_i=X_j]^{-1}\times(\tilde{\omega}_j-1)\mathbf{Z}_j(1-\delta_j)+o_p(1)$$

$$=-N^{-1/2}\sum_{j=1}^{N}(\tilde{\omega}_j-1)(1-\delta_j)\Pi_j+o_p(1),$$

where $\Pi_j=E[U_i\mathbf{Z}_i^{\mathsf{T}} \mid \tilde{\pi}_{0i}=\tilde{\pi}_{0j}, X_i=X_j]E[\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}} \mid \tilde{\pi}_{0i}=\tilde{\pi}_{0j}, X_i=X_j]^{-1}\mathbf{Z}_j$, which can be regarded as a linear (conditional) projection of U_j onto the space of $\mathbf{Z}_j$ under the inner product $\langle X_i, Y_i\rangle = E(X_iY_i)$. Also note that $E[U_i\mathbf{Z}_i^{\mathsf{T}} \mid \tilde{\pi}_{0i}=\tilde{\pi}_{0j}, X_i=X_j]E[\mathbf{Z}_i\mathbf{Z}_i^{\mathsf{T}} \mid \tilde{\pi}_{0i}=\tilde{\pi}_{0j}, X_i=X_j]^{-1}$ is the minimizer of

$$\frac{1}{N}\sum_{i=1}^{N}(U_i-\theta\mathbf{Z}_i)^2K_b((\tilde{\pi}_{0i}, X_i)-(\tilde{\pi}_{0j}, X_j))$$

with respect to θ. So $E[\mathbf{Z}_i(U_i - \Pi_i) \mid \widetilde{\pi}_{0i}, X_i] = 0$. Since the first component of $\mathbf{Z}_j$ is one, we have that $E[(U_i - \Pi_i) \mid \widetilde{\pi}_{0i}, X_i] = 0$. So $\widehat{U}$ can be rewritten as

$$\widehat{U} = N^{-1/2} \sum_{i=1}^{N} U_i + N^{-1/2} \sum_{i=1}^{N} (1 - \delta_i)(\widetilde{\omega}_i - 1)(U_i - \Pi_i) + o_p(1). \tag{A.3}$$

It follows from Cai and Zheng[1] that $\widehat{U}$ is asymptotically normal, with asymptotic variance

$$\begin{aligned} \Sigma_U &= E(U_i^2) + EN^{-1} \sum_{i=1}^{N} (1 - \delta_i)(\widetilde{\omega}_i - 1)^2 (U_i - \Pi_i)^2 + o_p(1) \\ &= E(U_i^2) + E[(\frac{1 - \widetilde{\pi}_{0i}}{\widetilde{\pi}_{0i}})(1 - \delta_i)(U_i - \Pi_i)^2] + o_p(1). \end{aligned}$$

Because the interaction term is

$$\begin{aligned} & E[N^{-1} \sum_{i \neq j}^{N} (\widetilde{\omega}_i - 1)(\widetilde{\omega}_j - 1)(U_i - \Pi_i)(U_j - \Pi_j)] \\ &= (N - 1) E \operatorname{Cov}\big(\widetilde{\omega}_i \{U_i - \Pi_i\}, \widetilde{\omega}_j \{U_j - \Pi_j\} \mid \mathscr{D}\big) \\ &= -m(N - 1)/N \int \eta(t, X_i, \delta_i) \eta(t, X_j, \delta_j) \frac{\mathrm{d}\Lambda_{NCC}(t)}{P(X \geq t)} = 0, \end{aligned}$$

where $\Lambda_{NCC}(t) = \int_0^t \mathrm{d}\boldsymbol{A}_{NCC}(u) / \boldsymbol{P}(X \geq u)$, $\boldsymbol{A}_{NCC}(t) = E\{N_i(t)\}$ and

$$\begin{aligned} & \eta(t, X_i, \delta_i) E[\{U_i - \Pi_i\} I(X_i \geq t)(1 - \widetilde{\pi}_{0i}) / \widetilde{\pi}_{0i}] \\ &= E(E[\{U_i - \Pi_i\} I(X_i \geq t)(1 - \widetilde{\pi}_{0i}) / \widetilde{\pi}_{0i} \mid \widetilde{\pi}_{0i}, X_i]) = 0 \end{aligned} \tag{A.4}$$

by the arguments before (A.3) and similar arguments to those of Samuelsen[16].

From Cai and Zheng[1], we know that the asymptotic variance of the TIPW estimator $\widehat{U} = N^{-1/2} \sum_{i=1}^{N} \widetilde{\omega}_i \boldsymbol{U}_i$ is

$$\begin{aligned} \Sigma^{TIPW} &= E(U_i^2) + E(U_i^2 \frac{1 - \widetilde{\pi}_{0i}}{\widetilde{\pi}_{0i}}) - m \int \eta_u(t, X_i, \delta_i)^2 \frac{\mathrm{d}\Lambda_{NCC}(t)}{P(X \geq t)} + o_p(1) \\ &= E(U_i^2 / \widetilde{\pi}_{0i}) - m \int \eta_u(t, X_i, \delta_i)^2 \frac{\mathrm{d}\Lambda_{NCC}(t)}{P(X \geq t)} + o_p(1), \end{aligned}$$

where $\eta_u(t, X_i, \delta_i) = E[U_i I(X_i \geq t)(1 - \widetilde{\pi}_{0i}) / \widetilde{\pi}_{0i}]$.

Comparing these two asymptotic variances, we have

$$\begin{aligned}\Sigma^{TIPW} - \Sigma_U &= E[(1-\delta_i)(\frac{1-\widetilde{\pi}_{0i}}{\widetilde{\pi}_{0i}})\{U_i^2 - (U_i - \Pi_i)^2\}] - m\int \eta_u(t, X_i, \delta_i)^2 \frac{\mathrm{d}\Lambda_{NCC}(t)}{P(X \geq t)} \\ &= E\{(1-\delta_i)(\frac{1-\widetilde{\pi}_{0i}}{\widetilde{\pi}_{0i}})\Pi_i^2\} - m\int \eta_u(t, X_i, \delta_i)^2 \frac{\mathrm{d}\Lambda_{NCC}(t)}{P(X \geq t)} \\ &= var\{N^{-1/2}\sum_{i=1}^{N}(1-\delta_i)(\widetilde{\omega}_i - 1)\Pi_i\} \geq 0,\end{aligned}$$

where the last equality holds similarly to (A.4). That is, $E[U_i I(X_i \geq t)(1-\widetilde{\pi}_{0i}) / \widetilde{\pi}_{0i}] = E[\Pi_i I(X_i \geq t)(1-\widetilde{\pi}_{0i}) / \widetilde{\pi}_{0i}]$. Therefore, the proposed AIPW estimators are more efficient than the true weight based TIPW estimators.

APPENDIX B.

Now we derive the specific forms of U_i in the general form (A.2) for all the related estimators of interest. Then the asymptotic variances of these estimators can be obtained using the results in Appendix A.

For $\hat{\boldsymbol{\gamma}}$, similarly to Cai and Zheng[1], we have that

$$N^{1/2}(\hat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) = N^{-1/2}\sum_{i=1}^{N}\widehat{\omega}_i U_{\bar{\boldsymbol{\gamma}}i} + o_p(1),$$

$$\text{where } U_{\bar{\boldsymbol{\gamma}}i} = D(\bar{\boldsymbol{\gamma}})^{-1}\int\{\mathbf{Y}_i - \frac{I^{(1)}(t)}{I^{(0)}(t)}\}\,\mathrm{d}M_i(t),$$

$$D(\bar{\boldsymbol{\gamma}}) = N^{-1}\sum_{i=1}^{N}\delta_i\left\{\frac{I^{(2)}(X_i)I^{(0)}(X_i) - I^{(1)}(X_i)^{\otimes 2}}{I^{(0)}(X_i)^{\otimes 2}}\right\},$$

$$I^{(k)}(t, \boldsymbol{\gamma}) = N^{-1}\sum_{i=1}^{N}\widehat{\omega}_i I(X_i \geq t)\exp(\mathbf{Y}_i^{\mathsf{T}}\boldsymbol{\gamma})\mathbf{Y}_i^k, k = 0, 1, 2,$$

$$I^{(k)}(t) = N^{-1}\sum_{i=1}^{N}\widehat{\omega}_i I(X_i \geq t)\exp(\mathbf{Y}_i^{\mathsf{T}}\bar{\boldsymbol{\gamma}})\mathbf{Y}_i^k, k = 0, 1, 2,$$

$$A_i(t) = \int_0^t I(X_i \geq u)\exp(\mathbf{Y}_i^{\mathsf{T}}\bar{\boldsymbol{\gamma}})\,\mathrm{d}\Lambda_0(u),$$

$$\text{and } M_i(t) = N_i(t) - A_i(t).$$

For $\widehat{\Lambda}(t \mid r)$, we have

$$
\begin{aligned}
&N^{1/2}\{\widehat{\Lambda}(t \mid r) - \Lambda(t \mid r)\} \\
&= N^{1/2} \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Y}_i - r)\, \mathrm{d}N_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Y}_i - r) I(X_i \geq u)} - N^{1/2}\Lambda(t \mid r) \\
&= N^{1/2} \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Y}_i - r)\, \mathrm{d}M_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Y}_i - r) I(X_i \geq u)} \\
&= N^{1/2} \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Y}_i - r)\, \mathrm{d}M_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\widehat{\boldsymbol{\gamma}}^\top \mathbf{Y}_i - r) I(X_i \geq u)} - N^{1/2} \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r)\, \mathrm{d}M_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) I(X_i \geq u)} + N^{1/2} \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r)\, \mathrm{d}M_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) I(X_i \geq u)} \\
&= \left| \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i \dot{K}_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) / h \mathbf{Y}_i\, \mathrm{d}M_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) I(X_i \geq u)} \right. \\
&\quad \left. - \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r)\, \mathrm{d}M_i(u) \{\sum_{i=1}^N \widehat{\omega}_i \dot{K}_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) / h \mathbf{Y}_i I(X_i \geq u)\}}{\{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) I(X_i \geq u)\}^2} \right| N^{1/2}(\widehat{\boldsymbol{\gamma}} - \bar{\boldsymbol{\gamma}}) \\
&\quad + N^{1/2} \int_0^t \frac{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r)\, \mathrm{d}M_i(u)}{\sum_{i=1}^N \widehat{\omega}_i K_h(\mathbf{Y}_i^\top \bar{\boldsymbol{\gamma}} - r) I(X_i \geq u)}
\end{aligned}
$$

.

So the U_i form in (A.2) for $\widehat{\Lambda}(t \mid r)$ is

$$
\begin{aligned}
U_{\Lambda i}(t \mid r) = U_{\bar{\boldsymbol{\gamma}} i}^\top \Bigg[& \int_0^t \frac{N^{-1} \sum_{j=1}^N \widehat{\omega}_j \dot{K}_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) / h \mathbf{Y}_j \mathrm{d}M_j(u)}{N^{-1} \sum_{j=1}^N \widehat{\omega}_j K_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) I(X_j \geq u)} \\
& - \int_0^t \frac{\sum_{j=1}^N \widehat{\omega}_j K_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) \mathrm{d}M_j(u) \{\sum_{l=1}^N \widehat{\omega}_l \dot{K}_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) / h \mathbf{Y}_l I(X_l \geq u)\}}{\{\sum_{j=1}^N \widehat{\omega}_j K_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) I(X_j \geq u)\}^2} \Bigg] \\
& + \int_0^t \frac{K_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) \mathrm{d}M_j(u)}{N^{-1} \sum_{j=1}^N \widehat{\omega}_j K_h(\mathbf{Y}_j^\top \bar{\boldsymbol{\gamma}} - \boldsymbol{r}) I(X_j \geq u)} .
\end{aligned}
$$

Recalling that $\widehat{S}(t \mid r) = \exp\{-\widehat{\Lambda}(t \mid r)\}$, we have that the U_i form in (A.2) for $\widehat{S}(t \mid r)$ is

$$
U_{Si}(t \mid r) = -S(t \mid r) U_{\Lambda i}(t \mid r) .
$$

Recalling that $\widehat{F}(r) = \dfrac{\sum_{i=1}^N \widehat{\omega}_i I(\widehat{\boldsymbol{R}}_{\text{all},\, i} \leq r)}{\sum_{i=1}^N \widehat{\omega}_i}$, we get that the U_i form in (A.2) for $\widehat{F}(r)$ is

$$
U_{Fi}(r) = I(\boldsymbol{R}_i \leq r) - F(c) + D_{\bar{\boldsymbol{\gamma}}}(r) U_{\bar{\boldsymbol{\gamma}} i}, \text{ where } D_{\bar{\boldsymbol{\gamma}}}(r) = \partial E[I(\boldsymbol{R}_i \leq r)] / \partial \boldsymbol{\gamma} \mid_{\boldsymbol{\gamma} = \bar{\boldsymbol{\gamma}}} .
$$

Recalling $\widehat{S}(r, t) = \int_r^\infty \widehat{S}(t \mid u)\, \mathrm{d}\widehat{F}(u)$, we have that the U_i form in (A.2) for $\widehat{S}(r, t)$ is

$$
U_{Si}(t, r) = \int_r^\infty U_{Si}(t \mid u)\, \mathrm{d}F(u) + \int_r^\infty S(t \mid u)\, \mathrm{d}U_{F_i}(u) .
$$

It follows that of U_i forms for the accuracy parameter estimators are

$$U_{TPR_t i}(r) = \frac{TPR_t(r)U_{Si}(t, r_l) - U_{Fi}(r) - U_{Si}(t, r)}{1 - S(t)},$$
$$U_{FPR_t i}(r) = \frac{U_{Si}(t, r) - FPR_t(r)U_{Si}(t, r_l)}{S(t)},$$
$$U_{PPV_t i}(r) = \frac{\{PPV_t(r) - 1\}U_{Fi}(r) - U_{Si}(t, r)}{1 - F(r)},$$
$$U_{NPV_t i}(r) = \frac{U_{Si}(t) - U_{Si}(t, r) - NPV_t(r)U_{Fi}(r)}{F(r)}.$$

Thus, we get the forms of U_i in (A.2) for the regression parameter estimator $\hat{\boldsymbol{\gamma}}$ and the accuracy parameter estimators $\widehat{\mathrm{TPR}}(c \mid t)$, $\widehat{\mathrm{FPR}}(c \mid t)$, $\widehat{\mathrm{PPV}}(c \mid t)$, $\widehat{\mathrm{NPV}}(c \mid t)$.

APPENDIX C.

In this section, we show the validity of the proposed resampling technique.

The derivative of $\widehat{\mathbf{U}}^*_{\widetilde{\pi}_{0i}, X_i}(\beta_i)$ with respect to β_i^* is

$$\frac{\partial \widehat{\mathbf{U}}^*_{\widetilde{\pi}_{0i}, X_i}(\beta_i^*)}{\partial \beta_i^*} = -\frac{1}{N}\sum_{j=1}^{N} I_j \frac{\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)}{\{1 + \exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)\}^2}\mathbf{Z}_j\mathbf{Z}_j^{\mathsf{T}}(1 - \delta_j)K_b((\widetilde{\pi}_{0i}, X_i) - (\widetilde{\pi}_{0j}, X_j)) + o_p(1),$$

which also converges to $-\Sigma_i$. It follows that

$$N^{1/2}(\hat{\beta}_i^* - \beta_i) = \Sigma_i^{-1}N^{-1/2}\sum_{j=1}^{N} I_j(V_j - \frac{\exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)}{1 + \exp(\beta_i^{\mathsf{T}}\mathbf{Z}_j)})\mathbf{Z}_j(1 - \delta_j)K_b((\widetilde{\pi}_{0i}, X_i) - (\widetilde{\pi}_{0j}, X_j)) + o_p(1).$$

The perturbed form of (A.2) is

$$\begin{aligned}
\hat{U}^* &= N^{-1/2}\sum_{i=1}^{N}[\delta_i I_i + (1 - \delta_i)V_{0i}I_i / \hat{\pi}_{0i}^*]U_i \\
&= N^{-1/2}\sum_{i=1}^{N}\{I_i\delta_i + (1 - \delta_i)I_i + (1 - \delta_i)(\frac{V_{0i}}{\widetilde{\pi}_{0i}} - 1)I_i + (1 - \delta_i)[\frac{V_{0i}I_i}{\hat{\pi}_{0i}^*} - \frac{V_{0i}I_i}{\widetilde{\pi}_{0i}}]\}U_i \\
&= N^{-1/2}\sum_{i=1}^{N} I_iU_i + N^{-1/2}\sum_{i=1}^{N}(1 - \delta_i)I_i(\frac{V_{0i}}{\widetilde{\pi}_{0i}} - 1)U_i - N^{-1/2}\sum_{i=1}^{N}(1 - \delta_i)\frac{V_{0i}I_i}{\widetilde{\pi}_{0i}}\frac{\hat{\pi}_{0i}^* - \widetilde{\pi}_{0i}}{\hat{\pi}_{0i}^*}U_i \\
&= N^{-1/2}\sum_{i=1}^{N} I_iU_i + N^{-1/2}\sum_{i=1}^{N}(1 - \delta_i)I_i(\widetilde{\omega}_i - 1)(U_i - \Pi_i) + o_p(1),
\end{aligned}$$

where the last equation follows similarly to the derivation of I_3 in Appendix A.

From (A.3), we know

$$\hat{U} = N^{-1/2}\sum_{i=1}^{N} U_i + N^{-1/2}\sum_{i=1}^{N}(1 - \delta_i)(\widetilde{\omega}_i - 1)(U_i - \Pi_i) + o_p(1).$$

It follows that

$$\widehat{U}^* - \widehat{U} = N^{-1/2} \sum_{i=1}^{N} (I_i - 1)U_i + N^{-1/2} \sum_{i=1}^{N} (1 - \delta_i)(I_i - 1)(\widetilde{\omega}_i - 1)(U_i - \Pi_i) + o_p(1).$$

Therefore,

$$\mathrm{Var}(\widehat{U}^* - \widehat{U} \mid \mathscr{D}) = \mathrm{Var}(\widehat{U}).$$

References

1. Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers under nested case-control studies. Biostatistics 2012; 13(1): 89–100. [PubMed: 21856652]
2. Zheng Y, Brown M, Lok A, Cai T, others. Improving efficiency in biomarker incremental value evaluation under two-phase designs. The Annals of Applied Statistics 2017; 11(2): 638–654. [PubMed: 28943991]
3. Pepe M, Feng Z, Janes H, Bossuyt P, Potter J. Pivotal evaluation of the accuracy of a biomarker used for classification or prediction: standards for study design. Journal of the National Cancer Institute 2008; 100(20): 1432–1438. [PubMed: 18840817]
4. Johnson SR, Anderson GL, Barad DH, Stefanick ML. The Women's Health Initiative: rationale, design and progress report. British Menopause Society Journal 1999; 5(4): 155–159.
5. Colditz GA, MANSON JE, HANKINSON SE. The Nurses' Health Study: 20-year contribution to the understanding of health among women. Journal of Women's Health 1997; 6(1): 49–62.
6. Prentice RL, Breslow N. Retrospective studies and failure time models. Biometrika 1978: 153–158.
7. Breslow NE, Day NE, others. Statistical Methods in Cancer Research. 1 International Agency for Research on Cancer Lyon. 1980.
8. Prentice R A case-cohort design for epidemiologic cohort studies and disease prevention trials. Biometrika 1986; 73(1): 1.
9. Martin LJ, Melnichouk O, Huszti E, et al. Serum lipids, lipoproteins, and risk of breast cancer: a nested case-control study using multiple time points. JNCI: Journal of the National Cancer Institute 2015; 107(5).
10. Chambers JC, Loh M, Lehne B, et al. Epigenome-wide association of DNA methylation markers in peripheral blood from Indian Asians and Europeans with incident type 2 diabetes: a nested case-control study. The lancet Diabetes & endocrinology 2015; 3(7): 526–534. [PubMed: 26095709]
11. 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(2): e000232.
12. Goldstein L, Langholz B. Asymptotic theory for nested case-control sampling in the Cox regression model. The Annals of Statistics 1992: 1903–1928.
13. Langholz B, Borgan Ø. Estimation of absolute risk from nested case-control data. Biometrics 1997: 767–774. [PubMed: 9192463]
14. Scheike TH, Juul A. Maximum likelihood estimation for Cox's regression model under nested case-control sampling. Biostatistics 2004; 5(2): 193–206. [PubMed: 15054025]
15. 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(3): 486–502. [PubMed: 16500923]
16. Samuelsen SO. A psudolikelihood approach to analysis of nested case-control studies. Biometrika 1997; 84(2): 379–394.
17. Lu W, Liu M. On estimation of linear transformation models with nested case–control sampling. Lifetime data analysis 2012; 18(1): 80–93. [PubMed: 21912975]

18. Cai T, Zheng Y. Evaluating prognostic accuracy of biomarkers in nested case-control studies. Biostatistics 2011; 13(1): 89–100. [PubMed: 21856652]
19. Cai T, Zheng Y. Nonparametric evaluation of biomarker accuracy under nested case-control studies. Journal of the American Statistical Association 2011; 106(494): 569–580. [PubMed: 22844169]
20. Lin DY, Wei LJ. The robust inference for the Cox proportional hazards model. Journal of the American statistical Association 1989; 84(408): 1074–1078.
21. Cai T, Tian L, Uno H, Solomon SD, Wei L. Calibrating parametric subject-specific risk estimation. Biometrika 2010; 97(2): 389–404. [PubMed: 23049123]
22. Davidson WS, Silva RGD, Chantepie S, Lagor WR, Chapman MJ, Kontush A. Proteomic analysis of defined HDL subpopulations reveals particle-specific protein clusters: relevance to antioxidative function. Arteriosclerosis, thrombosis, and vascular biology 2009; 29(6): 870–876.
23. Andrikoula M, McDowell I. The contribution of ApoB and ApoA1 measurements to cardiovascular risk assessment. Diabetes, Obesity and Metabolism 2008; 10(4): 271–278.
24. Movva R, Rader DJ. Laboratory assessment of HDL heterogeneity and function. Clinical Chemistry 2008; 54(5): 788–800. [PubMed: 18375481]
25. Kohan AB. ApoC-III: a potent modulator of hypertriglyceridemia and cardiovascular disease. Current opinion in endocrinology, diabetes, and obesity 2015; 22(2): 119.
26. Colditz GA, Philpott SE, Hankinson SE. The impact of the Nurses' Health Study on population health: prevention, translation, and control. American journal of public health 2016; 106(9): 1540–1545. [PubMed: 27459441]
27. Kang S Fitting semiparametric accelerated failure time models for nested case–control data. Journal of Statistical Computation and Simulation 2017; 87(4): 652–663.
28. Andersen PK, Gill RD. Cox's regression model for counting processes: a large sample study. The annals of statistics 1982: 1100–1120.
29. Du Y, Akritas M. Uniform strong representation of the conditional Kaplan-Meier process. Mathematical Methods of Statistics 2002; 11(2): 152–182.

TABLE 1

The Bias, empirical standard error (SE) and relative efficiency (RE) of the TIPW estimator, the proposed AIPW estimator, the nonparametric method based estimator (NP) and the CLR based estimator (CLR). For the proposed AIPW estimator, we also calculated the average of the estimated standard error (ASE), empirical coverage probabilities (CP) of the 95% CIs (×100) for settings (1).

Independent censoring (I)

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.010	0.022	0.005	0.146	0.096	0.096	0.096	0.166	1.000	0.955	1.020	0.191	0.090	93.0
$\bar{\gamma}_2$	1.099	−0.005	0.014	−0.022	0.344	0.100	0.095	0.092	0.340	1.000	1.086	1.116	0.043	0.090	93.6
TPR	0.460	−0.011	−0.010	−0.025		0.050	0.044	0.046		1.000	1.296	0.942		0.044	94.2
PPV	0.543	−0.012	−0.005	0.002		0.035	0.032	0.034		1.000	1.285	1.132		0.032	94.4
NPV	0.932	−0.001	−0.002	−0.008		0.011	0.007	0.009		1.000	2.220	0.863		0.008	96.8
TPR	0.596	−0.010	−0.009	−0.027		0.051	0.041	0.044		1.000	1.492	1.010		0.042	94.8
PPV	0.435	−0.009	−0.003	0.005		0.030	0.027	0.030		1.000	1.325	1.056		0.026	94.0
NPV	0.945	−0.001	−0.001	−0.008		0.011	0.007	0.008		1.000	2.586	1.007		0.007	97.0
TPR	0.748	−0.006	−0.005	−0.022		0.046	0.035	0.038		1.000	1.747	1.163		0.036	96.2
PPV	0.326	−0.005	−0.000	0.009		0.024	0.021	0.023		1.000	1.429	0.972		0.021	95.2
NPV	0.961	−0.001	−0.001	−0.006		0.011	0.006	0.007		1.000	2.986	1.229		0.007	97.2

Dependent censoring (II)

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.018	0.044	0.007	0.086	0.116	0.111	0.115	0.169	1.000	0.952	1.027	0.382	0.106	91.9
$\bar{\gamma}_2$	1.099	−0.004	−0.014	−0.019	0.210	0.122	0.105	0.101	0.315	1.000	1.319	1.398	0.104	0.101	93.3
TPR	0.460	−0.010	−0.005	−0.025		0.052	0.046	0.050		1.000	1.303	0.891		0.047	96.0
PPV	0.543	−0.010	−0.002	0.005		0.039	0.033	0.037		1.000	1.483	1.150		0.033	94.8
NPV	0.932	−0.001	−0.002	−0.009		0.010	0.008	0.009		1.000	1.623	0.575		0.008	96.8
TPR	0.596	−0.008	−0.004	−0.026		0.048	0.042	0.044		1.000	1.308	0.886		0.044	94.6
PPV	0.435	−0.006	−0.000	0.008		0.032	0.026	0.029		1.000	1.533	1.137		0.027	95.0
NPV	0.945	−0.001	−0.001	−0.008		0.009	0.007	0.009		1.000	1.578	0.606		0.008	96.0
TPR	0.748	−0.008	−0.004	−0.025		0.042	0.037	0.040		1.000	1.310	0.833		0.038	95.0
PPV	0.326	−0.005	0.000	0.010		0.025	0.021	0.024		1.000	1.560	0.987		0.021	95.6
NPV	0.961	−0.001	−0.001	−0.007		0.008	0.007	0.008		1.000	1.580	0.602		0.007	96.6

TABLE 2

The Bias, empirical standard error (SE) and relative efficiency (RE) of the TIPW estimator, the proposed AIPW estimator, the nonparametric method based estimator (NP) and the CLR based estimator (CLR). For the proposed AIPW estimator, we also calculated the average of the estimated standard error (ASE), empirical coverage probabilities (CP) of the 95% CIs (×100) for settings (2).

Independent censoring (I)

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.014	0.020	−0.009	0.139	0.110	0.105	0.118	0.171	1.000	1.078	0.880	0.254	0.096	91.9
$\bar{\gamma}_2$	1.099	−0.018	0.013	−0.097	0.343	0.109	0.097	0.145	0.339	1.000	1.263	0.399	0.052	0.096	94.0
TPR	0.460	−0.015	−0.006	−0.052		0.049	0.043	0.057		1.000	1.383	0.446		0.046	95.8
PPV	0.543	−0.003	−0.001	0.054		0.034	0.030	0.041		1.000	1.309	0.261		0.033	96.8
NPV	0.932	−0.004	−0.002	−0.036		0.011	0.008	0.017		1.000	2.124	0.085		0.009	97.2
TPR	0.596	−0.016	−0.006	−0.064		0.048	0.043	0.055		1.000	1.391	0.360		0.043	95.6
PPV	0.435	−0.001	0.000	0.057		0.029	0.025	0.040		1.000	1.302	0.176		0.027	96.6
NPV	0.945	−0.004	−0.002	−0.033		0.010	0.007	0.016		1.000	2.101	0.086		0.008	97.4
TPR	0.748	−0.013	−0.004	−0.072		0.045	0.039	0.051		1.000	1.480	0.287		0.039	94.6
PPV	0.326	0.002	0.002	0.056		0.025	0.021	0.035		1.000	1.444	0.142		0.022	95.6
NPV	0.961	−0.003	−0.001	−0.031		0.009	0.007	0.015		1.000	2.091	0.085		0.008	97.8

Dependent censoring (II)

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.024	0.044	−0.001	0.088	0.121	0.115	0.131	0.172	1.000	1.008	0.892	0.408	0.106	91.7
$\bar{\gamma}_2$	1.099	−0.009	−0.008	−0.098	0.230	0.130	0.110	0.133	0.331	1.000	1.389	0.622	0.104	0.102	93.6
TPR	0.460	−0.010	−0.002	−0.049		0.056	0.050	0.058		1.000	1.302	0.565		0.048	92.8
PPV	0.543	0.001	0.004	0.064		0.040	0.035	0.046		1.000	1.311	0.263		0.034	94.1
NPV	0.932	−0.004	−0.003	−0.038		0.011	0.008	0.017		1.000	2.181	0.085		0.009	95.6
TPR	0.596	−0.012	−0.003	−0.067		0.053	0.045	0.056		1.000	1.426	0.385		0.044	94.9
PPV	0.435	0.002	0.005	0.066		0.034	0.029	0.040		1.000	1.335	0.190		0.028	92.6
NPV	0.945	−0.004	−0.002	−0.036		0.011	0.007	0.016		1.000	2.360	0.080		0.008	97.9
TPR	0.748	−0.011	−0.000	−0.072		0.045	0.036	0.047		1.000	1.631	0.289		0.038	95.3
PPV	0.326	0.004	0.006	0.065		0.027	0.022	0.036		1.000	1.506	0.137		0.022	95.1
NPV	0.961	−0.003	−0.001	−0.033		0.010	0.006	0.014		1.000	2.594	0.079		0.007	97.7

TABLE 3

The Bias, empirical standard error (SE) and relative efficiency (RE) of the TIPW estimator, the proposed AIPW estimator, the nonparametric method based estimator (NP) and the CLR based estimator (CLR). For the proposed AIPW estimator, we also calculated the average of the estimated standard error (ASE), empirical coverage probabilities (CP) of the 95% CIs (×100) for settings (3).

Independent censoring (I)

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.028	0.024	0.007	0.152	0.129	0.101	0.099	0.169	1.000	1.614	1.758	0.335	0.105	94.6
$\bar{\gamma}_2$	1.099	−0.086	−0.002	−0.063	0.335	0.128	0.094	0.092	0.345	1.000	2.716	1.908	0.103	0.103	95.6
TPR	0.460	−0.039	−0.009	−0.036		0.053	0.048	0.050		1.000	1.777	1.130		0.050	94.0
PPV	0.543	0.015	−0.002	0.020		0.040	0.033	0.036		1.000	1.671	1.056		0.037	96.6
NPV	0.932	−0.017	−0.003	−0.017		0.014	0.008	0.011		1.000	6.487	1.131		0.010	98.0
TPR	0.596	−0.043	−0.006	−0.038		0.058	0.048	0.051		1.000	2.266	1.293		0.048	92.8
PPV	0.435	0.020	0.001	0.025		0.036	0.027	0.032		1.000	2.251	1.040		0.031	96.6
NPV	0.945	−0.015	−0.001	−0.015		0.014	0.008	0.011		1.000	6.411	1.182		0.009	96.2
TPR	0.748	−0.045	−0.001	−0.035		0.057	0.040	0.045		1.000	3.244	1.621		0.042	94.6
PPV	0.326	0.022	0.003	0.027		0.033	0.021	0.025		1.000	3.568	1.099		0.025	96.8
NPV	0.961	−0.014	−0.001	−0.013		0.013	0.007	0.010		1.000	6.815	1.307		0.008	96.2

Dependent censoring (II)

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.019	0.032	0.003	0.074	0.138	0.115	0.116	0.159	1.000	1.372	1.443	0.629	0.109	92.1
$\bar{\gamma}_2$	1.099	−0.082	−0.016	−0.061	0.213	0.129	0.106	0.104	0.354	1.000	2.032	1.603	0.136	0.102	93.9
TPR	0.460	−0.032	−0.001	−0.034		0.063	0.052	0.054		1.000	1.816	1.206		0.050	91.9
PPV	0.543	0.022	0.006	0.027		0.045	0.034	0.040		1.000	2.154	1.116		0.036	94.1
NPV	0.932	−0.017	−0.003	−0.019		0.014	0.009	0.012		1.000	5.339	1.012		0.010	97.0
TPR	0.596	−0.036	0.001	−0.035		0.061	0.049	0.049		1.000	2.134	1.427		0.047	93.1
PPV	0.435	0.026	0.008	0.032		0.039	0.028	0.033		1.000	2.692	1.042		0.030	94.5
NPV	0.945	−0.015	−0.002	−0.016		0.013	0.009	0.011		1.000	5.235	1.097		0.009	96.3
TPR	0.748	−0.038	0.003	−0.035		0.053	0.042	0.043		1.000	2.340	1.379		0.040	94.3
PPV	0.326	0.026	0.008	0.033		0.034	0.021	0.027		1.000	3.593	1.009		0.024	96.1
NPV	0.961	−0.014	−0.001	−0.014		0.012	0.008	0.010		1.000	4.634	1.091		0.008	97.6

TABLE 4

The Bias, empirical standard error (SE) and relative efficiency (RE) of the TIPW estimator, the proposed AIPW estimator, the nonparametric method based estimator (NP) and the CLR based estimator (CLR). For the proposed AIPW estimator, we also calculated the average of the estimated standard error (ASE), empirical coverage probabilities (CP) of the 95% CIs (×100).

		Bias				SE				RE				AIPW	
	true	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**TIPW**	**AIPW**	**NP**	**CLR**	**ASE**	**CP**
$\bar{\gamma}_1$	0.693	0.022	0.053	0.006	0.022	0.122	0.103	0.106	0.108	1.000	1.152	1.372	1.275	0.105	93.5
$\bar{\gamma}_2$	1.099	−0.062	−0.008	−0.027	0.043	0.116	0.087	0.085	0.193	1.000	2.245	2.146	0.440	0.093	94.9
TPR	0.457	−0.044	−0.009	−0.036		0.047	0.038	0.043		1.000	2.623	1.305		0.038	91.5
PPV	0.163	0.004	−0.002	0.002		0.022	0.016	0.020		1.000	1.866	1.291		0.016	94.9
NPV	0.988	−0.003	−0.000	−0.002		0.002	0.001	0.002		1.000	8.090	1.539		0.001	95.9
TPR	0.601	−0.049	−0.008	−0.039		0.045	0.034	0.038		1.000	3.471	1.465		0.036	93.3
PPV	0.113	0.005	−0.000	0.004		0.014	0.010	0.012		1.000	2.334	1.446		0.010	95.9
NPV	0.991	−0.003	−0.001	−0.002		0.002	0.001	0.001		1.000	8.275	1.503		0.001	95.1
TPR	0.757	−0.047	−0.005	−0.035		0.037	0.028	0.031		1.000	4.325	1.631		0.031	96.1
PPV	0.075	0.005	−0.001	0.003		0.008	0.006	0.007		1.000	2.804	1.482		0.006	96.7
NPV	0.994	−0.003	−0.001	−0.002		0.001	0.001	0.001		1.000	8.069	1.548		0.001	93.3

TABLE 5

Hazard ratio (HR) estimates for MI risk using sampling weights based on the original protocol (IPW), the proposed AIPW method and CLR method.

covariate	IPW (95% CI)	AIPW (95% CI)	CLR (95% CI)
smoking (past)	0.804 (0.151, 1.457)	1.054 (0.815, 1.363)	0.723 (0.431, 1.214)
smoking <15 cig/d	0.890 (0.240, 1.541)	1.222 (0.854, 1.748)	1.067 (0.838, 1.359)
smoking >15 cig/d	1.160 (0.847, 1.473)	1.253 (1.083, 1.449)	NA
age	1.014 (0.502, 1.526)	1.156 (0.852, 1.569)	0.379 (0.077, 1.863)
diabetes	1.573 (1.346, 1.800)	1.359 (1.121, 1.649)	1.335 (1.015, 1.757)
high cholesterol	1.380 (1.073, 1.686)	1.369 (1.073, 1.747)	1.349 (1.049, 1.736)
medication for HBP	1.307 (1.063, 1.550)	1.430 (1.193, 1.714)	1.301 (1.059, 1.599)
WPA1	0.600 (0.101, 1.098)	0.730 (0.500, 1.066)	0.688 (0.496, 0.953)
apoE	1.420 (1.107, 1.733)	1.427 (1.140, 1.786)	1.214 (0.950, 1.550)

TABLE 6

Estimated accuracy measures for a MI risk model with clinical predictors and biomarkers WPA1 and apoE and the incremental values (incV) of WPA1 and apoE over a model with only clinical predictors.

measure	FPR	est (95% CI)	incv (95% CI)
TPR	0.050	0.265 (0.146, 0.385)	−0.022 (−0.106, 0.061)
PPV	0.050	0.021 (0.009, 0.033)	−0.001 (−0.009, 0.007)
NPV	0.050	0.997 (0.996, 0.998)	0.000 (−0.000, 0.000)
TPR	0.100	0.360 (0.234, 0.487)	0.001 (−0.081, 0.083)
PPV	0.100	0.014 (0.008, 0.021)	−0.001 (−0.005, 0.004)
NPV	0.100	0.997 (0.996, 0.998)	0.000 (−0.000, 0.000)
TPR	0.200	0.503 (0.378, 0.627)	0.025 (−0.065, 0.115)
PPV	0.200	0.010 (0.007, 0.014)	0.000 (−0.002, 0.003)
NPV	0.200	0.998 (0.997, 0.998)	0.000 (−0.000, 0.001)
AUC		0.688 (0.610, 0.765)	−0.005 (−0.057, 0.047)