The Annals of Applied Statistics
2017, Vol. 11, No. 2, 638–654
DOI: 10.1214/16-AOAS997

IMPROVING EFFICIENCY IN BIOMARKER INCREMENTAL VALUE EVALUATION UNDER TWO-PHASE DESIGNS[1]

BY YINGYE ZHENG*, MARSHALL BROWN*, ANNA LOK† AND TIANXI CAI‡

Fred Hutchinson Cancer Research Center, University of Michigan† and Harvard T.H. Chan School of Public Health‡*

Cost-effective yet efficient designs are critical to the success of biomarker evaluation research. Two-phase sampling designs, under which expensive markers are only measured on a subsample of cases and noncases within a prospective cohort, are useful in novel biomarker studies for preserving study samples and minimizing cost of biomarker assaying. Statistical methods for quantifying the predictiveness of biomarkers under two-phase studies have been proposed [*Biostatistics* **13** (2012) 89–100, *Biometrics* **68** (2012) 1219–1227]. These methods are based on a class of inverse probability weighted (IPW) estimators where weights are "true" sampling weights that simply reflect the sampling strategy of the study. While simple to implement, existing IPW estimators are limited by lack of practicality and efficiency. In this manuscript, we investigate a variety of two-phase design options and provide statistical approaches aimed at improving the efficiency of simple IPW estimators by incorporating auxiliary information available for the entire cohort. We consider accuracy summary estimators that accommodate auxiliary information in the context of evaluating the incremental values of novel biomarkers over existing prediction tools. In addition, we evaluate the relative efficiency of a variety of sampling and estimation options under two-phase studies, shedding light on issues pertaining to both the design and analysis of biomarker validation studies. We apply our methods to the evaluation of a novel biomarker for liver cancer risk conducted with a two-phase nested case control design [*Gastroenterology* **138** (2010) 493–502].

1. Introduction.

Novel biomarkers have the potential to improve risk prediction for diseases such as cancer. Due to the cost associated with biomarker measurement, the improvement in the predictive performance of a model enriched with novel biomarkers over a model with only clinical risk factors, throughout referred to as the incremental value (IncV) of the novel biomarkers, needs to be rigorously assessed before incorporating the enriched risk model into routine clinical practice. A major barrier to validating prediction models is that measuring novel markers from a large prospective cohort study may be too expensive, especially if the event rate is low. Two subcohort sampling designs, the case cohort (CCH)

Received July 2016.

[1] Supported in part by Grants U01 CA86368, P01 CA053996, R01 GM085047, P50 MH106999 and P01 CA134294 awarded by the National Institutes of Health.

Key words and phrases. Biomarker, prediction accuracy, risk prediction, two-phase study.

[Prentice (1986)] and nested case control (NCC) [Thomas (1977)], are often employed as cost-effective alternatives to the standard full cohort design, and have been recently adopted for risk marker evaluation studies [Lok et al. (2010), Wang et al. (2011)].

These designs, while cost effective, can be challenging due to the outcome-dependent missingness on the marker information. Statistical methods have been developed to incorporate such missingness in estimating relative and absolute risk parameters [Borgan, Goldstein and Langholz (1995), Langholz and Borgan (1997), Self, Prentice et al. (1988)]. However, evaluation of the clinical utility of risk markers adds another level of complexity, requiring additional estimation of distribution of risks in the population and its summary indices. Appropriate statistical methods for risk model evaluation under two-phase studies and guidance to efficiently conduct the design are still lacking. Novel statistical tools that can be used for estimating the predictive performance of a single biomarker have also been developed for both CCH and NCC studies [Cai and Zheng (2012), Liu, Cai and Zheng (2012)]. In these approaches, simple inverse probability-weighted (IPW) estimators were considered, with weights as the reciprocals of true selection probabilities calculated based on the observed data and study design.

While such IPW estimators are simple to implement, limitations exist. First, in many practical situations, two-phase sampling plans can be quite complicated due to practical considerations such as the need to reuse samples previously assayed for other studies or missing measurements due to inadequate samples. Retrieving "true" sampling weights can therefore be considerably difficult in practice. In addition, these simple IPW estimators tend to be quite inefficient because they discard information from individuals without biomarker information. When auxiliary variables related to both outcome and incomplete marker measurement are available from the entire cohort, incorporating such information in estimation may lead to improvement in efficiency [Breslow et al. (2009a, 2009b), Saegusa and Wellner (2013)]. In this manuscript we propose novel estimators of prediction performance measures for two-phase studies, aiming to improve efficiency and practicality over existing estimators. Our estimators are based on the idea of augmentation, previously considered for estimating relative risk parameters under case-cohort studies. The augmented estimators adopt the IPW principal but use nonparametrically estimated weights based on auxiliary information.

Our second goal is to address study design issues, particularly regarding the impact of matching on estimation efficiency when the goal is to evaluate the IncV of novel biomarkers. In settings where routine markers or other auxiliary information exist, matching controls to cases on baseline predictors is usually considered. Matching is frequently adopted as a way to improve efficiency, particularly for the estimation of relative risk parameters [Breslow, Day et al. (1980)]. However, little is known regarding whether matching improves efficiency for the estimation of prediction performance and IncV measures. In addition, it has been previously noted that using augmented weights can lead to the efficiency gain of hazard ratio parameters for the fully observed covariates in a Cox regression model, but not so much for the partially observed biomarkers [Qi, Wang and Prentice (2005)]. The extent of efficiency gain due to augmentation for the estimation of prediction performance or IncV parameters has not yet been studied. In this paper, we perform extensive numerical studies to provide insight on the connection between the augmented estimators under minimally matched sampling designs and the simple IPW estimators under matched/stratified designs. We evaluate the relative efficiency of a variety of sampling and estimation options to identify strategies that are both efficient and practical.

2. Model specification and general estimation under two-phase studies.

2.1. *Notation.* Suppose the full cohort has N individuals from the targeted population followed prospectively. Due to censoring, the underlying full cohort data consist of N i.i.d. copies of the vector, $\mathscr{D} = \{\mathbf{D}_i = (X_i, \delta_i, \mathbf{Y}_i^{\mathsf{T}}, \mathbf{Z}_i^{\mathsf{T}})^{\mathsf{T}}, i = 1, \dots, N\}$, where $X_i = \min(T_i, C_i)$, $\delta_i = I(T_i \leq C_i)$, T_i and C_i denote failure time and censoring time, respectively, and subscript i indexes the subjects in the cohort. Here, $\mathbf{Y}_i = (\mathbf{Y}_{\mathrm{old}i}^{\mathsf{T}}, \mathbf{Y}_{\mathrm{new}i}^{\mathsf{T}})$ is the vector of all potential risk predictors, $\mathbf{Y}_{\mathrm{old}i}$ includes a set of routine markers available for all, $\mathbf{Y}_{\mathrm{new}i}$ represents novel risk markers only ascertained at the second phase for a selected subset of individuals, and $\mathbf{Z}_i$ represents auxiliary variables including matching and stratification variables available for the entire cohort. While $\mathbf{W}_i = (\mathbf{Z}_i^{\mathsf{T}}, \mathbf{Y}_{\mathrm{old}i})^{\mathsf{T}}$ is available for the entire cohort, $\mathbf{Y}_{\mathrm{new}i}$ is only available if $V_i = 1$, where V_i is a binary variable indicating whether subject i is selected to the phase II subcohort. The two-phase sampling only depends on X_i, δ_i and $\mathbf{Z}_i$, with the *true* sampling probability $\tilde{\pi}_i = P(V_i = 1|\mathscr{D}) = P(V_i = 1|X_i, \delta_i, \mathbf{Z}_i)$ known by design. We also assume that the risk $\mathcal{R}_t(\mathbf{Y}) = P(T_i \leq t|\mathbf{Y}_i = \mathbf{Y})$ follows a semiparametric transformation model [Cheng, Wei and Ying (1995, 1997), Zeng and Lin (2006)]

$$\begin{aligned} 1 - \mathcal{R}_t(\mathbf{Y}) &= \mathcal{T}_0[\log\{H(t)\} + \boldsymbol{\beta}_{\mathrm{new}}^{\mathsf{T}}\mathbf{Y}_{\mathrm{new}} + \boldsymbol{\beta}_{\mathrm{old}}^{\mathsf{T}}\mathbf{Y}_{\mathrm{old}}] \\ &= \mathcal{T}_0[\log\{H(t)\} + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{Y}], \end{aligned} \tag{2.1}$$

where $H(t)$ is an increasing function and $\mathcal{T}_0$ is a cumulative distributional function. When $\mathcal{T}_0(x) = \exp\{-\exp(x)\}$, the model corresponds to the proportional hazards model [Cox (1972)].

2.2. *A general inverse probability weighted framework for two-phase studies.* To incorporate outcome dependent missingness in $\mathbf{Y}_{\mathrm{new}}$, estimation of IPW procedures is based on subjects with $V_i = 1$ and reweights the ith observation by $\omega_i = V_i/\pi_i$. Consider a generic IPW statistic $\widehat{\mathbb{R}} = N^{-\frac{1}{2}} \sum_{i=1}^{N} \omega_i \mathbf{R}_i$, where $E(\mathbf{R}_i) = \mathbf{0}$. An obvious choice for π_i is the true sampling probability $\tilde{\pi}_i = P(V_i = 1|\mathscr{D})$, which leads to a class of True Weights based IPW (TIPW) statistics

$\widehat{\mathbb{R}}^{\text{TIPW}} = N^{-\frac{1}{2}} \sum_{i=1}^{N} \widetilde{\omega}_i \mathbf{R}_i$. The form of $\widetilde{\pi}_i$ can be obtained explicitly for both stratified CCH (sCCH) [Gray (2009), Liu, Cai and Zheng (2012)] and NCC [Cai and Zheng (2012), Samuelsen (1997)] designs; see Appendix A of the supplementary article [Zheng et al. (2017)] for details.

When $\widetilde{\pi}_i$ is not directly available from the study and/or to improve efficiency over the simple TIPW estimators, we focus on AIPW estimators that leverage information on auxiliary variables $\mathbf{W}$ by nonparametrically estimating π_i given $\mathbf{W}_i$. The AIPW approach replaces $\widetilde{\omega}_i$ with an augmented weight $\widehat{\omega}_i = V_i/\widehat{\pi}_i$, where $\widehat{\pi}_i = \widehat{\pi}(\mathbf{W}_i)$ is an estimate of $\widetilde{\pi}_i$ using $\mathbf{W}_i$. The key to the efficiency gain from the AIPW approach is to choose $\mathbf{W}$ and the estimator $\widehat{\pi}(\cdot)$ such that $E(\widehat{\omega}_i|\mathbf{W}_i) \approx 1$ and $\mathbf{W}$ is highly correlated with $\mathbf{R}_i$. For example, one may consider $\mathbf{W}_i^{\text{NCC}} = (\delta_i, X_i, \mathbf{Z}_i^{\mathsf{T}}, \mathbf{Y}_{\text{old}i}^{\mathsf{T}})^{\mathsf{T}}$ for mNCC design and $\mathbf{W}_i^{\text{CCH}} = (\delta_i, \mathbf{Z}_i^{\mathsf{T}}, \mathbf{Y}_{\text{old}i}^{\mathsf{T}})^{\mathsf{T}}$ for sCCH to enable both consistent estimation of the sampling weights and efficiency improvement by leveraging full cohort information on $Y_{\text{old}i}$.

When $\mathbf{W}_i$ is discrete, a natural choice for $\widehat{\pi}(\cdot)$ is the empirical proportion based on the observed data: $\widehat{\pi}(\mathbf{w}) = \frac{\sum_{i=1}^{N} V_i I(\mathbf{W}_i = \mathbf{w})}{\sum_{i=1}^{N} I(\mathbf{W}_i = \mathbf{w})}$. However, $\mathbf{W}_i$ often involves continuous variables. For example, for NCC designs, the sampling is dependent on X; thus, $\mathbf{W}$ needs to include X to ensure the consistency of the AIPW estimators. To incorporate continuous $\mathbf{W}$, one may consider the Nadaraya–Watson estimator,

$$\widehat{\pi}(\mathbf{w}) = \frac{\sum_{i=1}^{N} V_i K_h(\mathbf{w} - \mathbf{W}_i)}{\sum_{i=1}^{N} K_h(\mathbf{w} - \mathbf{W}_i)}, \tag{2.2}$$

where $K_h(\cdot) = K(\cdot/h)/h$, K is a symmetric kernel density function, and $h > 0$ is the bandwidth. Selection of appropriate h can follow the recommendations in Wang and Wang (2001) and Qi, Wang and Prentice (2005). Since the IPW estimators could be biased if $\widehat{\pi}_i$ does not consistently estimate $P(V_i = 1|\mathscr{D})$, such a widely applicable nonparametric estimator, applicable to a wide range of practical situations, is appealing.

Asymptotic behavior of AIPW estimators. Making inference under a two-phase design with weight is generally difficult because the sampling scheme leads to weak correlation between the V_i's, which is not ignorable even in large samples. Derivations for the asymptotic properties of the AIPW estimators accounting for such correlations are given in Appendix B of the supplementary article [Zheng et al. (2017)]. For ease of presentation, we focus on a setting where all cases are selected and controls are sampled according to the CCH or NCC design without additional matching. We also show in Appendix B that the variance reduction, $\mathbf{\Delta_R} = \text{Var}(\widehat{\mathbb{R}}^{\text{TIPW}}) - \text{Var}(\widehat{\mathbb{R}}^{\text{AIPW}})$, is always greater than or equal to 0, and thereby justifies the efficiency gain by AIPW estimators over the TIPW estimators.

3. Accuracy and incremental value evaluation.

3.1. *Parameters of interest.* For any subvector of $\mathbf{Y}$, $\mathbf{Y}_*$, and the associated risk model for $\mathcal{R}_t^*(\mathbf{Y}) \equiv P(D_t = 1|\mathbf{Y}_*)$, $\mathbf{Y}_*$ affects D_t only through the risk score $\mathcal{R}_t^*(\mathbf{Y})$. Thus, we quantify the predictiveness of $\mathbf{Y}_*$ based on the predictiveness of $\mathcal{R}_t^*(\mathbf{Y})$. One main goal here is to quantify the prediction performance of a risk score $\mathcal{R}_t^*(\mathbf{Y})$ for predicting $D_t = I(T \leq t)$ for various choices of $\mathcal{R}_t^*$. An array of measures can be considered for such evaluations. Key summary indices for characterizing the accuracy of $\mathcal{R}_t^*(\mathbf{Y})$ in classifying D_t include

$$\mathrm{TPR}_t^*(p) = P[\mathcal{R}_t^*(\mathbf{Y}) \geq p|T \leq t] \quad \text{and} \quad \mathrm{FPR}_t^*(p) = P[\mathcal{R}_t^*(\mathbf{Y}) \geq p|T > t],$$
$$\mathrm{PPV}_t^*(p) = P[T \leq t|\mathcal{R}_t^*(\mathbf{Y}) \geq p] \quad \text{and} \quad \mathrm{NPV}_t^*(p) = P[T > t|\mathcal{R}_t^*(\mathbf{Y}) < p],$$

where p is a risk threshold that can potentially be used to form different clinical decisions.

The pair of summaries $\mathrm{TPR}_t^*(p)$ and $\mathrm{FPR}_t^*(p)$ specifies the cumulative distribution of risks among t-year cases with $D_t = 1$ and noncases with $D_t = 0$, respectively, and is a building block for other measures. For example, taking D_t as a binary outcome for a fixed t, the proportion of t-year cases followed [Pfeiffer and Gail (2011)] can be expressed as $\mathrm{PCF}_t^*(v) = \mathrm{TPR}_t^*\{\mathcal{V}^{*-1}(1 - v)\}$, where $\mathcal{V}^*(p) \equiv \mathrm{P}\{\mathcal{R}_t^*(\mathbf{Y}) \leq p\}$. Its inverse function $\mathrm{PNF}_t^*(p) = \mathrm{PCF}_t^{*-1}(p)$ is the fraction of the general population at the highest risk that needs to be followed to ensure that a fraction p of the t-year cases will be captured.

When no specific risk thresholds are of key interest, one may consider summary measures to complement the display of case and control risk distributions. For example,

$$\mathrm{AUC}_t^* = \int \mathrm{TPR}_t^*\{\mathrm{FPR}_t^{*-1}(u)\}\,du = P\{\mathcal{R}_t^*(\mathbf{Y}_i) > \mathcal{R}_t^*(\mathbf{Y}_j)|T_i \leq t, T_j > t\}$$

is a time-dependent version of the area under the ROC curve (AUC), which provides a measure of separation between the distributions of $\mathcal{R}_t^*(\mathbf{Y})$ among t-year cases and noncases. Another frequently used prediction performance measure is the difference in mean risks (DMR) between cases and noncases at time t, which is related to the Integrated Discrimination Improvement (IDI) statistic for comparing risk models [Pencina et al. (2008)], $\mathrm{DMR}_t^* = E\{\mathcal{R}_t^*(\mathbf{Y})|T \leq t\} - E\{\mathcal{R}_t^*(\mathbf{Y})|T > t\}$.

To quantify the IncV in risk prediction based on a generic prediction summary index denoted by $\mathcal{A}_t$, one may consider $\mathrm{IncV}_{\mathcal{A}_t} = \mathcal{A}_t^{\mathrm{upd}} - \mathcal{A}_t^{\mathrm{old}}$, where $\mathcal{A}_t^{\mathrm{upd}}$ is evaluated for the updated model $\mathcal{R}_t(\mathbf{Y})$ constructed with $\mathbf{Y} = (\mathbf{Y}_{\mathrm{new}}^{\mathsf{T}}, \mathbf{Y}_{\mathrm{old}}^{\mathsf{T}})^{\mathsf{T}}$ as predictors, and $\mathcal{A}_t^{\mathrm{old}}$ is the corresponding value for the risk model $\mathcal{R}_t^{\mathrm{old}}(\mathbf{Y}) = P(D_t = 1|\mathbf{Y}_{\mathrm{old}})$ developed using only $\mathbf{Y}_{\mathrm{old}}$.

3.2. *Estimation and inference of accuracy summaries and IncV.* We now investigate the AIPW estimation procedures for the evaluation of IncV under the semiparametric transformation model as specified in (2.1). Specifically following the approaches taken in Murphy, Rossini and van der Vaart (1997) and Zeng and Lin (2006), the model parameters $\boldsymbol{\beta} = (\boldsymbol{\beta}_{\text{old}}^{\mathsf{T}}, \boldsymbol{\beta}_{\text{new}}^{\mathsf{T}})^{\mathsf{T}}$ can be obtained by maximizing a weighted semiparametric likelihood:

$$\begin{aligned}\widehat{\ell}(H, \boldsymbol{\beta}) = \sum_{i=1}^{N} \widehat{w}_i \big(\delta_i \big[\log \lambda_1\{e^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Y}_i} H(X_i)\} + \log \Delta H(X_i) + \boldsymbol{\beta}^{\mathsf{T}}\mathbf{Y}_i\big] \\ - \Lambda_1\{e^{\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Y}_i} H(X_i)\}\big),\end{aligned}$$

where $\Lambda_1(x) = -\log \mathcal{T}_1(x)$, $\mathcal{T}_1(x) = \mathcal{T}_0\{\log(x)\}$ and $\lambda_1(x) = d\Lambda_1(x)/dx$, $\Delta H(x) = H(x) - H(x-)$. With $\widehat{\boldsymbol{\beta}}$ as estimators for $\boldsymbol{\beta}$, we can calculate $\widehat{\mathcal{R}}_t(\mathbf{Y}) = 1 - \mathcal{T}_0[\log\{\widehat{H}(t)\} + \widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{Y}]$, where $\widehat{H}(t) = \widehat{H}(t; \widehat{\boldsymbol{\beta}})$.

To estimate the pair of key predictive performance summaries, $\mathrm{TPR}_t^*(p)$ and $\mathrm{FPR}_t^*(p)$, for a generic risk function $\mathcal{R}_t^*(\mathbf{Y})$, we first note that under model (2.1),

$$\mathrm{TPR}_t^*(p) = \frac{E\{I(\mathcal{R}_t^*(\mathbf{Y}) \geq p)\mathcal{R}_t(\mathbf{Y})\}}{E\{\mathcal{R}_t(\mathbf{Y})\}}$$

and

$$\mathrm{FPR}_t^*(p) = \frac{E[I(\mathcal{R}_t^*(\mathbf{Y}) \geq p)\{1 - \mathcal{R}_t(\mathbf{Y})\}]}{E\{1 - \mathcal{R}_t(\mathbf{Y})\}}.$$

We assume that (2.1) holds, but allow the risk function $\mathcal{R}_t^*(\mathbf{Y})$ to be derived from a potentially misspecified submodel. This along with the AIPW principle motivates us to estimate $\mathrm{TPR}_t^*(p)$ and $\mathrm{FPR}_t^*(p)$, respectively, as

$$\widehat{\mathrm{TPR}}_t^*(p) = \frac{\sum_{i=1}^{N} \widehat{\omega}_i \widehat{\mathcal{R}}_t(\mathbf{Y}_i) I\{\widehat{\mathcal{R}}_t^*(\mathbf{Y}_i) \geq p\}}{\sum_{i=1}^{N} \widehat{\omega}_i \widehat{\mathcal{R}}_t(\mathbf{Y}_i)}, \tag{3.1}$$

$$\widehat{\mathrm{FPR}}_t^*(p) = \frac{\sum_{i=1}^{N} \widehat{\omega}_i \{1 - \widehat{\mathcal{R}}_t(\mathbf{Y}_i)\} I\{\widehat{\mathcal{R}}_t^*(\mathbf{Y}_i) \geq p\}}{\sum_{i=1}^{N} \widehat{\omega}_i \{1 - \widehat{\mathcal{R}}_t(\mathbf{Y}_i)\}}, \tag{3.2}$$

where $\widehat{\mathcal{R}}_t^*(\mathbf{Y})$ is the estimated risk function derived under the submodel for $\mathcal{R}_t^*(\mathbf{Y}) = P(D_t = 1 | \mathbf{Y}_*)$. Subsequently, we may construct augmented estimators for other risk parameters. For example, we estimate $\mathrm{PCF}_t^*(v)$ as $\widehat{\mathrm{PCF}}_t^*(v) = \widehat{\mathrm{TPR}}_t^*\{\widehat{\mathcal{V}}^{*-1}(1 - v)\}$, and $\mathrm{PNF}_t^*(q)$ as $\widehat{\mathrm{PNF}}_t^*(q) = \widehat{\mathrm{PCF}}_t^{*-1}(q)$, where $\widehat{\mathcal{V}}^*(p) = N^{-1} \sum_{i=1}^{N} \widehat{w}_i I\{\widehat{\mathcal{R}}_t^*(\mathbf{Y}_i) \leq p\}$. An estimator for DMR_t^* is $\widehat{\mathrm{DMR}}_t^* = \widehat{\mathrm{ITP}}_t^* - \widehat{\mathrm{IFP}}_t^*$, with $\widehat{\mathrm{ITP}}_t^* = \int_p \widehat{\mathrm{TPR}}_t^*(p)\, dp$ and $\widehat{\mathrm{IFP}}_t^* = \int_p \widehat{\mathrm{FPR}}_t^*(p)\, dp$, and an estimator for AUC_t^* is $\widehat{\mathrm{AUC}}_t^* = \int_u \widehat{\mathrm{TPR}}_t^*(u) \widehat{\mathrm{FPR}}_t^{*-1}(du)$.

For IncV evaluations, we compare the prediction performance of the $\mathcal{R}_t(\mathbf{Y})$ to that of $\mathcal{R}_t^{\text{old}}(\mathbf{Y})$ obtained by fitting (2.1) with $\mathbf{Y}_{\text{old}}$ only. When the full cohort data are available for $\mathbf{Y}_{\mathrm{old}}$, the estimation of model parameters associated with $P(D_t = 1|\mathbf{Y}_{\mathrm{old}})$ can be obtained using the standard procedures as in Zeng and Lin (2006) without weighting.

For a generic prediction accuracy parameter $\mathcal{A}_t$ representing either $\mathrm{TPR}_t(p)$, $\mathrm{FPR}_t(p)$, $\mathrm{PCF}_t(v)$, $\mathrm{PNF}_t(p)$, AUC_t or DMR_t, let $\mathcal{A}_t^{\mathrm{upd}}$, $\mathcal{A}_t^{\mathrm{old}}$, $\widehat{\mathcal{A}}_t^{\mathrm{upd}}$ and $\widehat{\mathcal{A}}_t^{\mathrm{old}}$ denote the true and estimated accuracy for $\mathcal{R}_t(\mathbf{Y})$ and $\mathcal{R}_t^{\mathrm{old}}(\mathbf{Y})$, respectively. The IncV with respect to $\mathcal{A}_t$, $\mathrm{IncV}_{\mathcal{A}_t} = \mathcal{A}_t^{\mathrm{upd}} - \mathcal{A}_t^{\mathrm{old}}$, can be calculated as $\widehat{\mathrm{IncV}}_{\mathcal{A}_t} = \widehat{\mathcal{A}}_t^{\mathrm{upd}} - \widehat{\mathcal{A}}_t^{\mathrm{old}}$.

To construct confidence intervals for the accuracy and IncV parameters, in the supplementary article, Appendix C [Zheng et al. (2017)], we provide the asymptotic variances of $\widehat{\mathbb{R}}_{\mathcal{A}_t}^{\mathrm{IncV}}$ for the CCH and NCC design based on the asymptotic linear expansion of $\widehat{\mathbb{R}}_{\mathcal{A}_t}^{\mathrm{upd}}$ and $\widehat{\mathbb{R}}_{\mathcal{A}_t}^{\mathrm{IncV}}$.

4. Simulations. We conducted simulations to examine the finite sample performances of our proposed procedures under both two-phase designs and the impact of different sampling and analysis strategies on efficiency. With a cohort of size $N = 5000$, we first generated Y_{old} and Y_{new} from a zero-mean bivariate normal distribution with unit variances and correlation 0.8. The event time T was generated by conditioning on Y_{old} and Y_{new} from $P(T \le t|\mathbf{Y}) = \mathcal{T}_0\{\log(\alpha_0 t) + \beta_1 Y_{\mathrm{old}} + \beta_2 Y_{\mathrm{new}}\}$, with $\mathcal{T}_0(x) = 1 - \exp\{-\exp(x)\}$, where $\beta_1 = \log(3)$, $\beta_2 = \log(2)$ and α_0 was chosen to be (i) 0.1 for studying CCH designs, representing a moderate event rate scenario; and (ii) 0.01 for NCC designs, representing a rare case scenario. The censoring time C was taken to be the minimum of 2 and W, where W followed a gamma distribution, with a shape parameter of 2.5 and a rate parameter of 2. The event rate was about 20% under the setting for studying CCH designs, and 4% under the setting for studying NCC designs. These two full cohort data-generating mechanisms were used for all simulation settings, and a variety of sampling strategies were implemented to assemble the phase II data. For each sampling design and parameter of interest, we obtained two IPW estimators: one with true sampling weights, and one with the weights estimated by nonparametrically estimating $P(V_i = 1|\mathbf{W})$ as in equation (2.2).

4.1. *Finite sample performance of the proposed estimators.* We first assessed the validity of our proposed inference procedures in finite samples. For simplicity, no additional matching variables were used for sampling. For the CCH design, we randomly sampled $n_1 = 105$ cases from $\{i : \delta_i = 1\}$ and $n_0 = 3n_1$ controls from $\{i : \delta_i = 0\}$. For the NCC design, we included all individuals with $\delta = 1$ as cases, and, for each case, we randomly selected 3 controls from the risk set of the case. To estimate the sampling weights for augmentation, we let $\mathbf{W} = (\delta, Y_{\mathrm{old}})^{\mathsf{T}}$ for CCH and $\mathbf{W} = (\delta, X, Y_{\mathrm{old}})^{\mathsf{T}}$ for NCC.

Based on the results of 5000 simulated datasets as shown in the supplementary article, Table 1 [Zheng et al. (2017)], we found that all point estimates had negligible bias. The asymptotic-based standard error estimators approximated the empirical standard errors well with empirical coverage levels of the 95% confidence intervals close to the nominal level for all parameters except NPV under the NCC design. This was not surprising because, in this case, the true NPV levels were extremely close to 1, which made finite sample standard error and interval estimation generally difficult as in any binomial proportion estimation setting [Brown, Cai and DasGupta (2002)]. We also varied the values of bandwidth in the nonparametric kernels to evaluate the robustness of the proposed estimators. Varying bandwidths had little impact on estimates of accuracy summaries in the simulated settings as shown in the supplementary article, Table 2 [Zheng et al. (2017)]. Reducing cohort size to 1000 for CCH and 2000 for NCC showed efficiency improvement of AIPW estimators over TIPW estimators, with slightly increasing bias; see the supplementary article [Zheng et al. (2017)], Tables 3(a) and (b), for details.

4.2. *Relative efficiency of different sampling and analytical options*. We conducted simulation studies to examine the effect of matching or stratification by a discrete variable Z on the efficiency of estimating various accuracy summaries. We let $Z = \sum_{l=1}^{2} I(Y_{\text{old}} \leq y_{q_l})$, where y_q is the $100q$th percentile of Y_{old} and $\{q_1, q_2\}$ are chosen as (i) $\{0.5, 0.75\}$ for the CCH design and (ii) $\{0.33, 0.66\}$ for the NCC design. We compared the efficiency of AIPW and TIPW estimators obtained with data generated from different sampling designs, with and without matching on Z.

CCH design. Irrespective of sampling strategy, a total of 150 cases with $\delta = 1$ and 450 controls with $\delta = 0$ were included in the phase II subcohort. Three commonly adopted sampling strategies were considered:

- Setting A (random): randomly sampled 150 cases and 450 controls without considering Z.
- Setting B (frequency matched): randomly sampled 150 cases, then sampled controls such that the distribution of Z among the selected controls was the same as that of the cases.
- Setting C (balanced design): sampled 50 cases and 150 controls from each stratum defined by the level of Z. This design led to oversampling categories with lower prevalence.

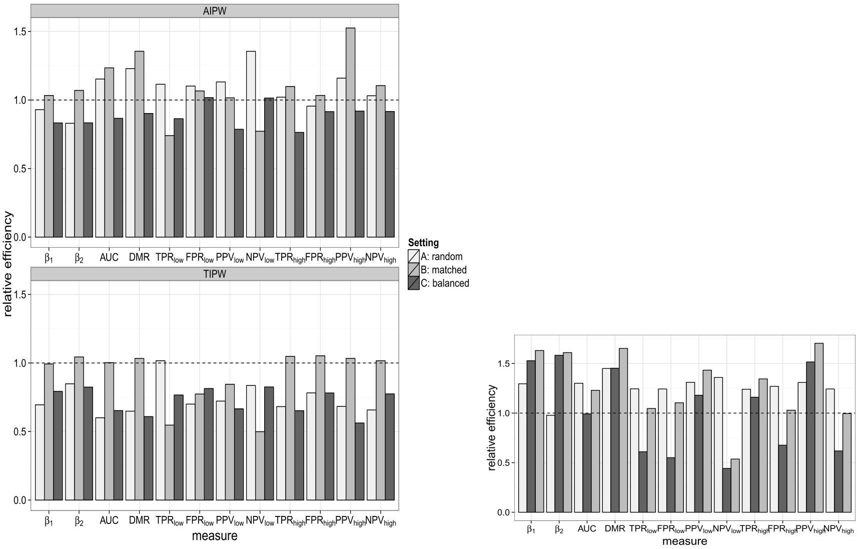
Under the CCH design, for any given parameter of interest $\mathcal{A}$ that is estimated via the TIPW approach as $\widehat{\mathcal{A}}$, it is possible to calculate the optimal sampling fractions to minimize the variance of $\widehat{\mathcal{A}}$ [Borgan et al. (2000)]. Suppose $\widehat{\mathcal{A}} - \mathcal{A} = N^{-1} \sum_{i=1}^{N} \widetilde{\omega}_i R_{\mathcal{A}i} + o_p(N^{-\frac{1}{2}})$, and the target is to sample n_1 cases and n_0 controls. Then the optimal sampling fractions that minimize the variance of $\widehat{\mathcal{A}}$ are π_{l1}^*

and π^*_{l0} for the cases and controls with $Z = l$, respectively, where

$$\pi^*_{ld} = \frac{n_d}{N_d} \frac{\mathrm{Var}(\mathbf{R}_{\mathcal{A}} | \delta = d, Z = l)^{1/2}}{\sum_{l'=1}^{L} v_{l'd} \mathrm{Var}(\mathbf{R}_{\mathcal{A}} | \delta = d, Z = j)^{1/2}},$$

$v_{l'd} = \sum_{i=1}^{N} I(\delta_i = d, Z_i = l')/N_d$ and $N_d = \sum_{i=1}^{N} I(\delta_i = d)$. Note that, in practice, an estimate of π^*_{ld} may only be available to assist in study design if preliminary data are available. In addition, such optimal sampling fractions tend to vary by specific measure—the sampling fractions optimal for one measure may not be optimal for the other. Thus, it is not possible to design a study to achieve optimal efficiency simultaneously for all measures. To mimic the most likely scenario in practice, we calculated optimal fractions for β_{new} and used them as the basis for sampling. The IPW estimators with true weights obtained under such a design (using sampling fraction optimal for β_{new}), denoted by TIPW$^{\text{opt}}$, were then used as the benchmark for comparing the efficiency of various standard designs and gauging the effect of augmentation.

Figure 1(a) shows the efficiencies of the TIPW and AIPW estimators obtained under various sampling strategies, relative to the TIPW$^{\text{opt}}$ estimators. When true



(a) CCH (b) NCC

FIG. 1. *Relative efficiency* (*RE*) *of various predictive performance summaries by different designs. Figure* (a): *results for CCH designs. Setting A*: *simple random sampling*; *Setting B*: *matched design*; *and Setting C*: *balanced design. TIPW* (*top panel*) *and AIPW* (*bottom pane*) *estimators for each setting are considered using TIPW*$^{\text{opt}}$ *as a benchmark for efficiency. Figure* (b): *results for NCC designs. Setting D*: *simple random sample*; *Setting E*: *matched design. TIPW estimator under setting D is the benchmark for efficiency.*

weights were used, the frequency matched design had similar efficiency as the optimal design and outperformed the random and balanced designs for a majority of the parameters investigated. However, for the accuracy parameters at various risk threshold levels, the efficiency of the frequency matching was much lower than the optimal design and was comparable to or sometimes worse than the random and balanced designs. On the other hand, the AIPW estimators were substantially more efficient than their corresponding TIPW estimators, except for β_2. Interestingly, the AIPW estimators under random sampling had efficiency comparable to or higher than those obtained from TIPW$^{\text{opt}}$ for all parameters of interest. When comparing the AIPW estimators obtained across the three designs, the balanced design generally performed the worst. Although the frequency matched design achieves a slightly higher efficiency for a few parameters than the random design, the random design appeared to be much more robust with regard to efficiency of AIPW estimators across different parameters. The results here suggested that, in practice, considering a simple random sampling scheme at the design stage and then utilizing auxiliary information in the analysis step has the advantage in both practical simplicity and statistical efficiency.

NCC design. For the NCC design, all cases were included and 3 controls were sampled from the risk sets of the cases according to the following two strategies:

- Setting D (random): randomly sampled from the risk set of the case.
- Setting E (matched): randomly sampled from the case's risk set and matched on the value of Z of the case.

Since no simple optimal sampling strategies can be implemented for the NCC design, we used the TIPW estimator under random sampling as the benchmark for comparison, and present in Figure 1(b) the efficiencies of the TIPW and AIPW estimators obtained under these two designs relative to the benchmark estimator. The matched design led to the most efficient relative risk estimators for β_1 and β_2, however, the efficiency gain did not directly translate to the estimation of performance summary parameters, and it may in fact lead to poorer efficiency compared to a simple random sampling design. Indeed, for a majority of summary performance parameters considered, Setting D, using a simple random sampling design with the proposed AIPW estimators, appeared to be the most efficient. The results further suggested the benefit of employing a simple random design followed by the AIPW estimation procedure. Stratifying/matching based on $\mathbf{Y}_{\text{old}}$, while leading to improved efficiency for the regression parameters, could drastically sacrifice the efficiency for various accuracy parameters. On the other hand, the AIPW estimator with random sampling always resulted in efficiency improvement. Numerical results are presented in the supplementary article [Zheng et al. (2017)], Tables 4(a) and (b).

5. Example.

Patients with hepatocellular carcinoma (HCC) often have poor prognosis due to late diagnosis. Since cirrhosis of any cause and chronic infection with hepatitis B virus (HBV) or hepatitis C virus (HCV) are the most common risk factors for HCC, surveillance of high-risk populations may detect tumors at an early stage when curative interventions can be implemented. Alpha fetoprotein (AFP) is the most widely used biomarker for HCC surveillance; however, its sensitivity and specificity in detecting early HCC are low. More reliable biomarkers for HCC surveillance and early detection are sought in order to improve the outcome of the disease.

The Hepatitis C Antiviral Long-Term Treatment against Cirrhosis (HALT-C) Trial included 1050 patients with chronic hepatitis C and bridging fibrosis or cirrhosis who failed to achieve a sustained virologic response (SVR) to a combination therapy of pegylated interferon and ribavirin. Patients were randomized to low-dose pegylated interferon or no treatment and examined every 3 months for a total duration of 3.5 years. Blood samples were collected at each visit for subsequent research testing, including assays for HCC biomarkers. Ultrasound examinations were repeated 6 months after enrollment and again every 12 months. Patients with an elevated or rising AFP and those with new lesions detected by ultrasound were evaluated further by CT or MRI.

One goal of the HALT-C Trial was to identify and validate markers for HCC surveillance. As part of the trial, an NCC study was employed to assess and compare the accuracy of AFP and a novel serum biomarker, des-gamma-carboxy prothrombin (DCP), in predicting the risk of HCC. The NCC subcohort included all 39 HCC cases diagnosed during the follow-up. For each case, 2 controls without HCC, matched for treatment assignment and presence of cirrhosis on baseline biopsy, were selected from the risk set of the case. This resulted in a total of 77 controls in the NCC subcohort. The biomarkers were evaluated at multiple follow-up visits, and the results on the repeated markers were reported in Lok et al. (2010), where conditional logistic regression models were used to compare characteristics of HCC cases, and matched controls and unconditional logistic regression were used to evaluate the accuracy performance of the biomarkers.

To illustrate our proposed methods, only baseline measurements were considered for risk modeling. Logarithm transformed values were considered for both AFP and DCP, denoted by logAFP and logDCP, respectively. Due to low liver cancer incidence, methods that could improve efficiency would be helpful. For comparison, we obtained parameter estimates using both the TIPW and AIPW approaches, where for the AIPW approach we let $\mathbf{W} = (X, \delta, \log \mathrm{AFP})^{\mathsf{T}}$ for augmentation. To build a risk model with both logAFP and logDCP, we considered fitting a Cox proportional hazards model. We obtained log hazard ratio (logHR) parameter estimates with the conditional logistic regression, TIPW, and AIPW methods. The conditional logistic regression method yielded a logHR estimate of 0.54 with a standard error (SE) of 0.27 for logAFP and 1.54 with a SE of 0.51 for logDCP, suggesting that DCP may serve as an independent risk factor for HCC beyond AFP.

The logHR was estimated as 0.61 (SE: 0.22) for logAFP and 2.04 (SE: 0.33) for logDCP based on TIPW, and 0.82 (SE: 0.18) for logAFP and 1.95 (SE: 0.32) for logDCP based on AIPW. These results indicated that the AIPW method provided more efficient estimates of the logHR parameters when compared to TIPW and conditional logistic regression methods.

We subsequently evaluated the 2-year predictive performance by combining logDCP and logAFP using the measures described in Section 3.1. The results for evaluating the full model with both logAFP and logDCP included are presented in the first two columns of Table 1. Across the measures we considered, point estimates from the two approaches in general were quite close; however, the AIPW estimators had substantially smaller standard errors than that of the TIPW estimators for most of the parameters. Combing AFP and DCP led to a good predictive model for predicting the 2-year risk of HCC, with AUC estimated as 0.81 (95%CI: [0.68, 0.94]) based on TIPW, and 0.82 (95%CI: [0.75, 0.90]) based on AIPW. If the top 20% of the population based on the estimated risks is considered of high risks, then the proportion of individuals who will be diagnosed with HCC within two years, captured by the prediction rule, is 64% (95% CI: [38%, 89%]) based on the TIPW estimate, and 68% (95% CI: [53%, 81%]) based on the AIPW estimate.

To further evaluate whether adding DCP to the model substantially improves accuracy when compared to the model with AFP alone, we also fit a model with AFP alone and calculated the IncV of DCP with respect to various accuracy parameters as shown in Table 1. The ROC curves and risk distribution for both models are shown Figure 2. As seen in the figures, the enriched model always had higher TPF and higher PCF, but smaller PNF across different risk thresholds p. Formal tests of such observed incremental values for selected p can be based on the results presented in the last two columns in Table 1. For example, $\mathrm{IncV}_{\mathrm{AUC}_2}$ was estimated as 12.9% with 95% CI (7.0%, 18.7%) based on AIPW, indicating that adding DCP improved in prediction accuracy beyond AFP. In addition, there was also significant improvement with respect to PCF and PNF, with $\mathrm{IncV}_{\mathrm{PCF}_2(0.2)}$ estimated as 21.4% (95% CI: [10.4%, 32.3%]) and $\mathrm{IncV}_{\mathrm{PNF}_2(0.2)}$ estimated as 18.3% (95% CI: [5.1%, 31.4%]), based on AIPW. The TIPW approach, while generating similar point estimates, did not produce statistically significant IncV estimates for all parameters considered (Table 1). This example demonstrates the advantage of the proposed AIPW method for estimating accuracy summaries and IncV parameters, particularly when there are limited samples with available biomarker measurements.

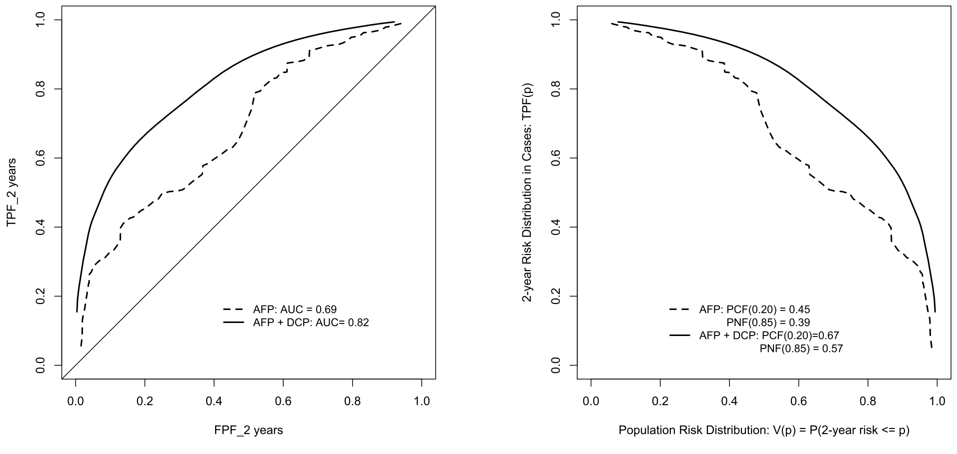
6. Discussion.

Large cohort biomarker studies of rare diseases such as cancer require thoughtful planning, from selection of study subjects and measurement of key variables and auxiliary information to analytical strategies. Study design becomes even more demanding in biomarker research when measurements are based on stored tissue or blood specimens. It is important in this setting to use research

TABLE 1

Evaluation of biomarkers in predicting 2*-year liver cancer incidence in Halt-C study. Shown below are TIPW and AIPW estimates along with their corresponding standard errors shown in parenthesis. The log-hazard ratios and accuracy parameters were estimated for* 2*-year risks estimated based on both* (i) *the full model with* $\log \mathrm{AFP} + \log \mathrm{DCP}$; *and* (ii) *the reduced model with* $\log \mathrm{AFP}$ *alone. Also presented are the TIPW and AIPW estimates for IncV parameters of* $\log \mathrm{DCP}$ *above and beyond* $\log \mathrm{AFP}$

	$\log \mathrm{AFP} + \log \mathrm{DCP}$		**$\log \mathrm{AFP}$**		**IncV of $\log \mathrm{DCP}$**	
Parameters	**TIPW (SE)**	**AIPW (SE)**	**TIPW (SE)**	**AIPW (SE)**	**TIPW (SE)**	**AIPW (SE)**
β_1	0.614 (0.221)	0.822 (0.177)	0.542 (0.206)	0.771 (0.168)	–	–
β_2	2.043 (0.333)	1.953 (0.316)	–	–	–	–
AUC	0.809 (0.066)	0.823 (0.040)	0.655 (0.079)	0.694 (0.054)	0.155 (0.129)	0.129 (0.030)
DMR	0.155 (0.376)	0.173 (0.193)	0.065 (0.148)	0.112 (0.081)	0.090 (0.473)	0.061 (0.267)
ITPR	0.168 (0.389)	0.188 (0.199)	0.108 (0.036)	0.142 (0.015)	0.060 (0.353)	0.045 (0.193)
IFPR	0.013 (0.013)	0.015 (0.007)	0.043 (0.154)	0.030 (0.087)	−0.029 (0.126)	−0.015 (0.082)
TPR(0.002)	0.979 (0.023)	0.981 (0.024)	0.986 (0.031)	0.989 (0.021)	−0.007 (0.031)	−0.008 (0.023)
FPR(0.002)	0.827 (0.123)	0.824 (0.167)	0.928 (0.060)	0.939 (0.066)	−0.101 (0.073)	−0.116 (0.125)
PPV(0.002)	0.014 (0.004)	0.016 (0.004)	0.016 (0.007)	0.014 (0.009)	−0.002 (0.007)	0.002 (0.007)
NPV(0.002)	0.999 (0.001)	0.999 (0.001)	0.997 (0.005)	0.998 (0.005)	0.001 (0.006)	0.001 (0.006)
TPR(0.02)	0.529 (0.287)	0.562 (0.083)	0.201 (0.098)	0.328 (0.050)	0.328 (0.292)	0.234 (0.069)
FPR(0.02)	0.112 (0.070)	0.114 (0.051)	0.069 (0.328)	0.099 (0.074)	0.043 (0.249)	0.015 (0.097)
PPV(0.02)	0.052 (0.015)	0.063 (0.019)	0.042 (0.041)	0.043 (0.041)	0.011 (0.038)	0.021 (0.055)
NPV(0.02)	0.994 (0.003)	0.993 (0.002)	0.987 (0.018)	0.990 (0.006)	0.007 (0.014)	0.003 (0.005)
PCF(0.20)	0.636 (0.131)	0.667 (0.071)	0.307 (0.047)	0.453 (0.048)	0.329 (0.117)	0.214 (0.056)
PNF(0.85)	0.526 (0.119)	0.570 (0.061)	0.400 (0.051)	0.386 (0.037)	0.125 (0.120)	0.183 (0.067)



(a) (b)

FIG. 2. *Comparing performance of two prediction models*: *model with AFP alone* (*solid lines*) *and model with both AFP and DCP* (*dashed lines*). (a) *ROC curves for predicting* 2-*year risk of HCC with baseline biomarker measurements*. (b): *risk distribution curves for individuals who were diagnosed with HCC in* 2 *years*.

resources wisely to achieve optimal efficiency of the study. There is a paucity of appropriate statistical methods for biomarker assessment and guidance on design and analysis strategies to maximize efficiency. Practical and efficient statistical tools can enable clinical investigators to conduct more cost-effective studies and more efficiently allocate research resources.

This manuscript contributes to such an endeavor in two ways. First, we provide a general framework for more efficiently estimating prediction accuracy and IncV parameters via an AIPW approach under two-phase CCH or NCC designs. Our simulation studies and application of Halt-C biomarker validation studies indicate that the use of nonparametric weights to capture design selection is valid and yields significant efficiency gain. Furthermore, the proposed approach also provides a practical solution in study settings where the true design-based sampling probabilities are impractical to ascertain. In addition, previous work on biomarker evaluation with two-phase studies [Cai and Zheng (2012), Liu, Cai and Zheng (2012)] only considered evaluating the performance of a single marker. We extend the scope of work to the evaluation of multivariate risk models and IncV of novel biomarkers under two-phase designs. Such extensions are nontrivial due to the complex structure induced by both the correlation among different risk markers and the sampling design.

Second, using extensive numerical studies, we demonstrated that stratification sampling for CCH studies or matching for NCC studies can be inefficient in many accuracy summaries when not done optimally. In the absence of preliminary data, it is often unclear what variables for matching or what sampling fractions should be considered for stratification. A poor choice in matching variables may lead to loss in efficiency and unnecessary complications in analysis. Furthermore, the sampling fractions optimal for one parameter may not be optimal for another, and thus no sampling strategies would be uniformly optimal across all parameters. Therefore, using a simple sampling scheme at the design stage and then improving estimation efficiency using the proposed augmented estimators in analysis would be a useful alternative to considering matched designs.

We have focused on the estimation of accuracy summaries with a semiparametric approach to illustrate the AIPW approach. Alternatively, one may consider calculating the accuracy summaries with a nonparametric approach as was previously considered [Cai and Zheng (2011)], without relying on the assumption of Model (2.1). The estimating and inference procedures with AIPW described can be easily adopted to that setting. Our proposed estimators for evaluating the IncV of a new prediction model improves efficiency of the existing IPW-based estimators; however, they do not achieve full efficiency as compared with a full likelihood-based approach [Zeng and Lin (2014)]. Future exploration of the additional gain when applying nonparametric likelihood-based procedures is warranted, even at the cost of increased computational burden. Finally, the validity of the class of IPW estimators is based on the assumption that selection is dependent on variables observable from the full cohort. Caution should be taken when the availability of biomarker measurement might be dependent on unmeasured variables.

Acknowledgments. We thank the Editor and anonymous reviewers whose careful guidance led to much improvement of the manuscript.

SUPPLEMENTARY MATERIAL

Supplementary Article for "Improving efficiency in biomarker incremental value evaluation under two-phase designs" (DOI: 10.1214/16-AOAS997SUPP; .pdf). We provide theoretical derivations and additional simulation results.

REFERENCES

BORGAN, Ø., GOLDSTEIN, L. and LANGHOLZ, B. (1995). Methods for the analysis of sampled cohort data in the Cox proportional hazards model. *Ann. Statist.* **23** 1749–1778. MR1370306

BORGAN, Ø., LANGHOLZ, B., SAMUELSEN, S. O., GOLDSTEIN, L. and POGODA, J. (2000). Exposure stratified case-cohort designs. *Lifetime Data Anal.* **6** 39–58. MR1767493

BRESLOW, N. E., DAY, N. E. et al. (1980). *Statistical Methods in Cancer Research*, *Vol.* 1. IARC Publications.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009a). Using the whole cohort in the analysis of case-cohort data. *Am. J. Epidemiol.* **169** 1398–1405.

BRESLOW, N. E., LUMLEY, T., BALLANTYNE, C. M., CHAMBLESS, L. E. and KULICH, M. (2009b). Improved Horvitz–Thompson estimation of model parameters from two-phase stratified samples: Applications in epidemiology. *Statistics in Biosciences* **1** 32–49.

BROWN, L. D., CAI, T. T. and DASGUPTA, A. (2002). Confidence intervals for a binomial proportion and asymptotic expansions. *Ann. Statist.* **30** 160–201. MR1892660

CAI, T. and ZHENG, Y. (2011). Nonparametric evaluation of biomarker accuracy under nested case-control studies. *J. Amer. Statist. Assoc.* **106** 569–580. MR2847971

CAI, T. and ZHENG, Y. (2012). Evaluating prognostic accuracy of biomarkers under nested case-control studies. *Biostatistics* **13** 89–100.

CHENG, S. C., WEI, L. J. and YING, Z. (1995). Analysis of transformation models with censored data. *Biometrika* **82** 835–845.

CHENG, S. C., WEI, L. J. and YING, Z. (1997). Predicting survival probabilities with semiparametric transformation models. *J. Amer. Statist. Assoc.* **92** 227–235. MR1436111

COX, D. R. (1972). Regression models and life-tables. *J. R. Stat. Soc., B* **34** 187–220. MR0341758

GRAY, R. J. (2009). Weighted analyses for cohort sampling designs. *Lifetime Data Anal.* **15** 24–40.

LANGHOLZ, B. and BORGAN, Y. (1997). Estimation of absolute risk from nested case-control data. *Biometrics* **53** 767–774.

LIU, D., CAI, T. and ZHENG, Y. (2012). Evaluating the predictive value of biomarkers with stratified case-cohort design. *Biometrics* **68** 1219–1227.

LOK, A. S., STERLING, R. K., EVERHART, J. E., WRIGHT, E. C., HOEFS, J. C., DI BISCEGLIE, A. M., MORGAN, T. R., KIM, H.-Y., LEE, W. M., BONKOVSKY, H. L. et al. (2010). Des-γ-carboxy prothrombin and α-fetoprotein as biomarkers for the early detection of hepatocellular carcinoma. *Gastroenterology* **138** 493–502.

MURPHY, S. A., ROSSINI, A. J. and VAN DER VAART, A. W. (1997). Maximum likelihood estimation in the proportional odds model. *J. Amer. Statist. Assoc.* **92** 968–976. MR1482127

PENCINA, M. J., D'AGOSTINO, R. B. SR., D'AGOSTINO, R. B. JR. and VASAN, R. S. (2008). Evaluating the added predictive ability of a new marker: From area under the ROC curve to reclassification and beyond. *Stat. Med.* **27** 157–172. MR2412695

PFEIFFER, R. M. and GAIL, M. H. (2011). Two criteria for evaluating risk prediction models. *Biometrics* **67** 1057–1065. MR2829240

PRENTICE, R. L. (1986). A case-cohort design for epidemiologic cohort studies and disease prevention trials. *Biometrika* **73** 1–11.

QI, L., WANG, C. Y. and PRENTICE, R. L. (2005). Weighted estimators for proportional hazards regression with missing covariates. *J. Amer. Statist. Assoc.* **100** 1250–1263. MR2236439

SAEGUSA, T. and WELLNER, J. A. (2013). Weighted likelihood estimation under two-phase sampling. *Ann. Statist.* **41** 269–295. MR3059418

SAMUELSEN, S. O. (1997). A pseudolikelihood approach to analysis of nested case-control studies. *Biometrika* **84** 379–394.

SELF, S. G., PRENTICE, R. L. (1988). Asymptotic distribution theory and efficiency results for case-cohort studies. *Ann. Statist.* **16** 64–81. MR0924857

THOMAS, D. C. (1977). Addendum to "Methods of cohort analysis: Appraisal by application to asbestos mining." *J. Roy. Statist. Soc. Ser. A* **140** 483–485.

WANG, S. and WANG, C. Y. (2001). A note on kernel assisted estimators in missing covariate regression. *Statist. Probab. Lett.* **55** 439–449. MR1877649

WANG, T., ROHAN, T. E., GUNTER, M. J., XUE, X., WACTAWSKI-WENDE, J., RAJPATHAK, S. N., CUSHMAN, M., STRICKLER, H. D., KAPLAN, R. C., WASSERTHEIL-SMOLLER, S., SCHERER, P. E. and GLORIA, Y. F. HO (2011). A prospective study of inflammation markers and endometrial cancer risk in postmenopausal hormone nonusers. *Cancer Epidemiol. Biomark. Prev.* **20** 971–977.

ZENG, D. and LIN, D. Y. (2006). Efficient estimation of semiparametric transformation models for counting processes. *Biometrika* **93** 627–640. MR2261447

ZENG, D. and LIN, D. Y. (2014). Efficient estimation of semiparametric transformation models for two-phase cohort studies. *J. Amer. Statist. Assoc.* **109** 371–383. MR3180570

ZHENG, Y., BROWN, M., LOK, A. and CAI, T. (2017). Supplement to "Improving Efficiency in Biomarker Incremental Value Evaluation under Two-phase Designs." DOI:10.1214/16-AOAS997SUPP.

Y. ZHENG
M. BROWN
FRED HUTCHINSON CANCER RESEARCH CENTER
1100 FAIRVIEW AVE. N.
SEATTLE, WASHINGTON 98109
USA
E-MAIL: yzheng@fhcrc.org
mdbrown@fredhutch.org

A. LOK
DEPARTMENT OF INTERNAL MEDICINE
UNIVERSITY OF MICHIGAN
1500 E. MEDICAL CENTER DR.
ANN ARBOR, MICHIGAN 48109
USA
E-MAIL: aslok@med.umich.edu

T. CAI
DEPARTMENT OF BIOSTATISTICS
HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH
655 HUNTINGTON AVE.
BOSTON MASSACHUSETTS 02115
USA
E-MAIL: tcai@hsph.harvard.edu