HHS Public Access

Author manuscript

Stat Med. Author manuscript; available in PMC 2020 January 15.

Published in final edited form as:

Stat Med. 2019 January 15; 38(1): 100–114. doi:10.1002/sim.7966.

Evaluating classification performance of biomarkers in two-phase case-control studies†

Lu Wang[1] and **Ying Huang**[*,1,2]

[1]Vaccine and Infectious Disease Division, Fred Hutchinson Cancer Research Center, Seattle, WA 98109, U.S.A.

[2]Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.

Summary

Biomarkers are playing an increasingly important role in disease screening, early detection, and risk prediction. The two-phase case-control sampling study design is widely used for the evaluation of candidate biomarkers. The sampling probabilities for cases and controls in the second phase can often depend on other covariates (sampling strata). This biased sampling can lead to invalid inference on a biomarker's classification accuracy if not properly accounted for. In this papeqer, we adopt the idea of inverse probability weighting (IPW) and develop IPW-based estimators for various measures of a biomarker's classification performance, including the points on the receiver operating characteristics (ROC) curve, the area under the ROC curve (AUC), and the partial AUC. In particular, we consider classification accuracy estimators using sampling weights estimated conditionally on sampling strata and further improve their efficiency through the use of estimated weights that additionally take into account the auxiliary variables available from the phase-one cohort. We develop asymptotic properties of the proposed estimators and provide analytical variance for making inference. Extensive simulation studies demonstrate excellent performance of the proposed weighted estimators, while the traditional empirical estimator can be severely biased. We also investigate the advantages in efficiency gain for estimating various classification accuracy estimators through the use of auxiliary variables in addition to sampling strata and apply the proposed method to examples from a renal artery stenosis study and a prostate cancer study.

Keywords

Inverse probability weighting; ROC curve; AUC; Partial AUC; Estimated weights; Two-phase sampling

†Evaluating classification performance of biomarkers in two-phase case-control studies.

[*]**Correspondence** Ying Huang, Fred Hutchinson Cancer Research Center, M2-C200, Seattle, WA 98109, U.S.A. yhuang@fhcrc.org.
Present Address Fred Hutchinson Cancer Research Center, M2-C200, Seattle, WA 98109, U.S.A.

SUPPORTING INFORMATION

Additional supplementary material may be found online in the supporting information tab for this article.

1 | INTRODUCTION

Biomarkers are playing an increasingly important role in disease screening, early detection, and risk prediction. A variety of statistical tools have been developed for characterizing the classification performance of biomarkers. The most commonly used tool for quantifying and visualizing a biomarker's classification accuracy for a binary disease outcome is the receiver operating characteristic (ROC) curve [1,2]. The area under the ROC curve (AUC) is a widely used summary measure for biomarker classification performance. This value corresponds to the probability that a randomly selected individual with the target disease has a marker value greater than a randomly selected individual free of the disease. In practice, a specific region under the ROC curve is often of greater interest than the entire region under the curve. For example, high sensitivity is desired for biomarkers used for disease diagnosis, while high specificity is essential for biomarkers used for screening purposes. The partial area under the ROC curve [3,4] has been proposed to characterize classification performance of biomarkers over a clinically relevant region.

In this paper, we evaluate biomarkers with respect to the points on the ROC curve, the AUC, and the partial AUC using data from two-phase sampling designs. In the first phase of these designs, a large cohort is randomly selected from the target population. For all subjects in the phase-one cohort, information is collected on the disease outcome of interest and on easily measured covariates. In the second phase, a subsample of subjects is drawn from the phase-one cohort without replacement and the biomarker is measured in this subsample. Here, we are particularly interested in case-control sampling scenarios in which cases and controls are separately drawn in the second phase, which has become a pervasive practice in biomarker studies [5].

Oftentimes in the second phase, cases and/or controls are not obtained by simple random sampling. For example, in some studies, cases and controls are randomly selected within pre-determined covariate strata; in others, a random sample of cases is obtained and then controls are selected to be frequency-matched to cases within covariate strata. These biased sampling schemes may impact biomarker evaluation if not properly accounted for. [6] demonstrated that when controls that are frequency-matched to cases on risk factors for the disease outcome are selected, estimates of biomarker performance can be seriously biased and lead to invalid inference. To overcome the problem that sampling probability for cases and/or controls varies across sampling strata, [7] adopted the idea of inverse probability weighting (IPW) and proposed IPW estimators for characterizing and comparing biomarker performance based on the AUC, where sampling weights were estimated conditional on the sampling strata. The proposed IPW AUC estimator was shown to be unbiased and the corresponding confidence interval based on analytically estimated variance had good coverage. Using analytical and numerical studies, [7] also demonstrated the efficiency advantage in the AUC estimator using an estimated sampling weight conditional on sampling strata even when the true sampling weight is known. In this paper, we extend the work of [7] and develop IPW estimators for the other two important measures of classification accuracy: the points on the ROC curve and the partial AUC. More importantly, to further improve the efficiency of IPW estimators, we consider the use of auxiliary variables correlated with the biomarker in estimation of sampling weights, in addition to the

use of sampling strata. In particular, we adjust the sampling weights via estimated weights conditional on sampling strata and additional auxiliary variables [8]. Estimated weights have been previously shown to improve the efficiency of weighted likelihood estimators in etiological studies [9,10]. We develop the asymptotic properties of the various IPW estimators for the points on the ROC curve, the AUC, and the partial AUC, and provide analytical variance formulas for making inference.

Two types of sampling designs, Bernoulli sampling [11] and finite-population stratified sampling [12], are considered in the second phase of the two-phase studies. In Bernoulli sampling, each subject has a pre-specified sampling probability and is drawn independently from the others in the second phase. In contrast, finite-population stratified sampling does not require pre-specified sampling probability of each subject in phase two, but the number of subjects sampled from each stratum is fixed. We demonstrate the connection in theoretical results of the proposed estimators between these two sampling designs.

This paper is organized as follows. In Section 2, we first propose the IPW estimators of the points on the ROC curve, the AUC, and the partial AUC in the Bernoulli sampling setting. We also establish the asymptotic properties of the proposed estimators. Then we consider the estimation in finite-population stratified sampling setting and describe the relationship between the asymptotic results of the proposed estimators in these two sampling designs. In Section 3, we conduct extensive simulation studies to evaluate the performance of the proposed estimators and assess the efficiency using different weight choices. In Section 4, we apply our proposed method to two clinical examples. Finally, we make concluding remarks in Section 5.

2 | METHODOLOGY AND THEORY

We consider a binary disease outcome D, with values 1 and 0 standing for diseased and non-diseased, respectively. Let X indicate a biomarker measured on a continuous scale. We consider a two-phase sampling setting for evaluating the classification performance of the biomarker. In the first phase of the two-phase sampling design, suppose there are N subjects randomly sampled from the target population, from whom the disease status D and a vector of easy-to-measure covariates Z are measured. The list of covariates Z measured in the first phase can include continuous variables that are discretized to form sampling strata for the second phase sampling, as well as additional auxiliary variables. Let N_D and $N_{\bar{D}}$ denote the numbers of cases and controls in the first phase, respectively ($N = N_D + N_{\bar{D}}$). Suppose that cases and controls among phase-one samples are classified into K_D and $K_{\bar{D}}$ sampling strata, with N_{Dk_D} and $N_{\bar{D}k_{\bar{D}}}$ samples in each stratum, respectively. In this case, $N_D = N_{D1} + N_{D2} + \cdots + N_{DK_D}$ and $N_{\bar{D}} = N_{\bar{D}1} + N_{\bar{D}2} + \cdots + N_{\bar{D}K_{\bar{D}}}$. In the second phase, subsamples of cases and controls are randomly drawn from the phase-one cohort within each sampling stratum and the continuous biomarker X is measured.

Next we propose estimators for (1) the points on the ROC curve, (2) the AUC, and (3) the partial AUC for biomarker X in two-phase case-control sampling studies. Each of these measures can be interpreted as a specific instance of the integrated ROC curve as pointed out by [13]. We start with the Bernoulli sampling design.

2.1 | Bernoulli Sampling

In Bernoulli sampling, subjects are selected at the second phase independently of one another, with a pre-specified sampling probability that is constant within each sampling stratum. Let δ_{Di} be the indicator of being sampled in phase-two for the i^{th} case and p_{Di} be the corresponding sampling probability. Similarly, let δ_{Dj} and p_{Dj} be the indicators of being sampled in the second phase for the j^{th} control and the corresponding sampling probability, respectively. We next consider the estimation of various classification performance measures with data collected from a Bernoulli sampling design.

2.1.1 | Estimation of points on the ROC Curve—First, we consider the estimation of points on the ROC curve. Let the cumulative distribution functions (CDFs) of the biomarker X among the cases and controls be F_D and $F_{\bar{D}}$, respectively. Suppose a binary test is constructed based on the biomarker with a threshold c such that individuals whose X value is greater than c are classified as diseased and all individuals whose X value is equal to or less than c are classified as non-diseased. The sensitivity of the test is defined as the probability of correctly classifying a diseased individual, i.e., $SEN(c) = 1 - F_D(c)$. The specificity of the test is defined as $SPE(c) = F_{\bar{D}}(c)$, i.e., the probability of correctly classifying a non-diseased individual. The ROC curve is created by plotting $SEN(c)$ against $1 - SPE(c)$ as the threshold c varies from $-\infty$ to $+\infty$. Consequently, the points on the ROC curve are $ROC(t) = 1 - F_D\left[F_{\bar{D}}^{-1}(1-t)\right]$, with $t \in [0, 1]$.

Based on the phase-two case/control sample, an empirical estimator of $ROC(t)$ can be constructed as $\widehat{ROC}_{em}(t) = 1 - \hat{F}_{D,em}\left[\hat{F}_{\bar{D},em}^{-1}(1-t)\right]$, where

$\hat{F}_{\bar{D},em}^{-1}(1-t) = \inf\left\{x : \hat{F}_{\bar{D},em}(x) \geq 1-t\right\}$ is the empirical estimator of $F_{\bar{D}}^{-1}(1-t)$,

$\hat{F}_{D,em}(x) = \left[\sum_{i=1}^{N_D} \delta_{Di} I(X_{Di} \leq x)\right] \Big/ \left(\sum_{i=1}^{N_D} \delta_{Di}\right)$ and

$\hat{F}_{\bar{D},em}(x) = \left[\sum_{j=1}^{N_{\bar{D}}} \delta_{\bar{D}j} I\left(X_{\bar{D}j} \leq x\right)\right] \Big/ \left(\sum_{j=1}^{N_{\bar{D}}} \delta_{\bar{D}j}\right)$ are the empirical estimators of F_D and $F_{\bar{D}}$, respectively. When case and/or control samples in the second phase are not representative of the corresponding case or control distributions in the population, the empirical estimator $\widehat{ROC}_{em}(t)$ could be seriously biased. To overcome this problem, we propose the following weighted estimator of $ROC(t)$ adopting the idea of inverse probability weighting (IPW) [14],

$$\widehat{ROC}_{IPW}(t) = 1 - \hat{F}_{D,IPW}\left[\hat{F}_{\bar{D},IPW}^{-1}(1-t)\right], \quad (1)$$

Where $\hat{F}_{D,IPW}(x) = \left[\sum_{i=1}^{N_D} \frac{\delta_{Di}}{\hat{p}_{Di}} I(X_{Di} \leq x)\right] \Big/ \left(\sum_{i=1}^{N_D} \frac{\delta_{Di}}{\hat{p}_{Di}}\right)$ and

$\hat{F}_{\bar{D},IPW}(x) = \left[\sum_{i=1}^{N_{\bar{D}}} \frac{\delta_{\bar{D}j}}{\hat{p}_{\bar{D}j}} I(X_{\bar{D}j} \leq x)\right] \Big/ \left(\sum_{j=1}^{N_{\bar{D}}} \frac{\delta_{\bar{D}j}}{\hat{p}_{\bar{D}j}}\right)$ are the IPW estimators of F_D and $F_{\bar{D}}$,

respectively, in which $\hat{p}_{Di}$ and $\hat{p}_{\bar{D}j}$ are the estimated sampling prabaoilities of case i and control j, respectively.

Note that in Bernoulli sampling design, p_{Di} and $p_{\bar{D}j}$, the sampling probabilities for cases and controls, are pre-specified and could be used for generating the weights in the weighted ROC estimator. However, using estimated sampling probability has been recommended as a way to improve estimation efficiency in many other settings such as weighted likelihood estimators [8, 9]. Moreover, in the paradigm of biomarker evaluation, estimating AUC using an estimated weight conditional on sampling stratum was demonstrated to improve efficiency [7]. Therefore, we focus on estimated sampling weights throughout this paper and propose different procedures to estimate the sampling weights below.

We consider two types of estimated sampling probabilities: i) the estimate conditional on the sampling stratum only, and ii) the estimate based on a model conditional on the whole list of covariates including sampling stratum and additional auxiliary variables available in the first phase. These two types of sampling probability estimators for cases/controls are denoted by $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$ and $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$, respectively. The estimated sampling weights are the inverse values of the corresponding estimated probabilities. We denote the corresponding IPW estimators of $ROC(t)$ as $\widehat{ROC}_{IPW}^{Str}(t)$ and $\widehat{ROC}_{IPW}^{Aux}(t)$.

First, since the sampling probability in phase two varies within some pre-specified strata, a natural estimator for the sampling probability of a case or control is the proportion of the phase-one cases or controls sampled in the second phase within the corresponding sampling stratum. For case i in stratum k_D, an empirical estimation of sampling probability within the stratum is $\hat{p}_{Di}^{Str} = n_{Dk_D} \Big/ N_{Dk_D}$, where n_{Dk_D} is the number of cases selected in the second phase within stratum k_D. Similarly, for control j in stratum $k_{\bar{D}}$, we have $\hat{p}_{\bar{D}j}^{Str} = n_{\bar{D}k_{\bar{D}}} \Big/ N_{\bar{D}k_{\bar{D}}}$, where $n_{\bar{D}k_{\bar{D}}}$ is the number of controls drawn in phase-two within stratum $k_{\bar{D}}$. This estimator of sampling weight was also adopted in [7] for constructing weighted AUC estimators.

More generally, we may have more information from phase one beyond the discrete sampling strata. Thus we endeavored to further improve the efficiency of the weighted estimator $\widehat{ROC}_{IPW}(t)$ by incorporating auxiliary variable information. The second type of sampling probability estimates we adopt are based on models of sampling probability conditional on sampling strata and available auxiliary variables in phase-one. Let V_D and $V_{\bar{D}}$ be the vectors of dummy variables indicating sampling strata with lengths $K_D - 1$ and $K_{\bar{D}} - 1$, respectively. Let Z_D and $Z_{\bar{D}}$ be the vectors of additional auxiliary variables for the cases and controls, respectively. Recall that Z_D and $Z_{\bar{D}}$ could include continuous variables (denoted as V_D^* and $V_{\bar{D}}^*$) that are discretized to form the sampling strata V_D and $V_{\bar{D}}$, as well as additional covariates measured in phase one. For example, in the Renal Artery Stenosis Study example presented later in Section 4, V_D^* and $V_{\bar{D}}^*$ are ages of cases and controls measured on a continuous scale in phase one, V_D and $V_{\bar{D}}$ are discretized age subgroups serving as sampling strata for phase-two sampling, and W are additional covariates

measured in phase one including sex, smoking history etc. We can model the probability of sampling cases/controls in phase two as a function of all available phase-one variables $V_D/V_{\bar{D}}$ and $Z_D/Z_{\bar{D}}$ using a generalized linear model (GLM), such as the logistic regression model [8]. That is, we model $p_D = g(\theta_D, V_D, Z_D)$ and $p_{\bar{D}} = g(\theta_{\bar{D}}, V_{\bar{D}}, Z_{\bar{D}})$ for some pre-specific function g, where θ_D and $\theta_{\bar{D}}$ are finite-dimensional parameters. For example, we can fit a logistic regression model for the sampling probabilities for cases/controls as

$$\begin{aligned} \log\left(\frac{p_D}{1-p_D}\right) &= \theta_{D0} + \theta_{D1}^T V_D + \theta_{D2}^T Z_D, \\ \log\left(\frac{p_{\bar{D}}}{1-p_{\bar{D}}}\right) &= \theta_{\bar{D}0} + \theta_{\bar{D}1}^T V_{\bar{D}} + \theta_{\bar{D}2}^T Z_{\bar{D}}, \end{aligned} \tag{2}$$

and obtain $\hat{p}_D^{Aux}$ and $\hat{p}_{\bar{D}}^{Aux}$ as the maximum likelihood estimators of p_D and $p_{\bar{D}}$ based on MLEs of θ_D and $\theta_{\bar{D}}$, respectively. Note that when the auxiliary variables Z_D and $Z_{\bar{D}}$ are not included in model (2) and the model is based on sampling strata V_D and $V_{\bar{D}}$ alone, the resulting estimates $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ turn into $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$. That is, the true model of sampling probability where sampling probability depends on sampling strata alone is nested within the more complicated model that includes additional covariates Z. Both models give consistent estimates of sampling probability.

In general, suppose we model the sampling probabilities for cases and controls with finite-dimensional parameters θ_D and $\theta_{\bar{D}}$, respectively. Let $\hat{\theta}_D$ and $\hat{\theta}_{\bar{D}}$ be the maximum likelihood estimators and $\hat{p}_D$ and $\hat{p}_{\bar{D}}$ be the corresponding sampling probabilities estimators. The corresponding IPW estimator of the points on the ROC curve (1) is asymptotically normally distributed, as stated below in Theorem 1.

Theorem 1. *Assume* $0 < p_D, p_{\bar{D}} \leq 1$ and $N_D/N \to \lambda \in (0, 1)$ *as the sample size* $N \to \infty$. *Then as* $N \to \infty$, $\sqrt{N}\left[\widehat{ROC}_{IPW}^{Aux}(t) - ROC(t)\right]$ *converges to a normal random variable with mean* 0 *and variance*

$$\sigma^2_{ROC(t),IPW} = \frac{1}{\lambda}\sigma^2_{D,IPW^{(c)}} + \frac{1}{1-\lambda}\frac{f^2_{\bar{D}}(c)}{f^2_{\bar{D}}(c)}\sigma^2_{\bar{D},IPW^{(c)}},$$

where

$$\begin{aligned} \sigma^2_{d,IPW^{(c)}} = {} & Var(H_d(c)) + E\left[\left(p_d^{-1} - 1\right)H_d(c)\right] - SS_d(c)\left\{E\left[\left(p_d^{-1} - 1\right)H_d(c)\right] + Cov\left(p_d^{-1} - 1, H_d(c)\right)\right\} \\ & -[SS_d(c) \times a_d - b_d(c)]^T I_d^{-1}[SS_d(c) \times a_d - b_d(c)], \text{ for } d = D, \bar{D} \end{aligned}$$

in which $c = F_{\bar{D}}^{-1}(1-t)$, $H_D(c) = I(X_D > c)$, $H_{\bar{D}}(c) = I(X_{\bar{D}} < c)$, $SS_D(c) = SEN(c)$, $SS_{\bar{D}}(c) = SPE(c)$, $a_d = E[(1/p_d)(\partial p_d/\partial\theta_d)]$, $b_d(c) = E[H_d(1/p_d)(\partial p_d/\partial\theta_d)]$, $I_d = E\left[\left(\frac{1}{p_d} + \frac{1}{1-p_d}\right)\frac{\partial p_d}{\partial\theta_d}\frac{\partial p_d}{\partial\theta_d}\right]$ *is the information matrix of* θ_d, *for* $d = D, \bar{D}$.

Proof of Theorem 1 is provided in supplementary material Appendix A. We note that $\sigma^2_{D,IPW}$ and $\sigma^2_{\bar{D},IPW}$ given in Theorem 1 are variance components due to variability of cases and controls, respectively. In fact, $\sigma^2_{D,IPW}$ and $\sigma^2_{\bar{D},IPW}$ equal the asymptotic variances of the IPW sensitivity and specificity estimators at threshold c, respectively. The corresponding result is illustrated in supplementary material Appendix A Lemma A.1. In further check of decomposition of $\sigma^2_{d,IPW}$, we observe that all parts except $V(H_d)$ are attributed to either the variability of sampling probability across strata or the estimation of sampling probability. In a special case with $p_D = p_{\bar{D}} = 1$, i.e.,for unbiased empirical $ROC(t)$ estimator, $\sigma^2_{d,IPW}$ is reduced to be $V_{ar}(H_d)$ for $d = D, \bar{D}$, and the corresponding $\sigma^2_{ROC(t),IPW}$ is equivalent to the asymptotic variance of the empirical $ROC(t)$ estimator derived from Theorem 2.2 in [15]. The component $E\left[\left(p_d^{-1}-1\right)H_d(c)\right] - SS_d(c)\left\{E\left[\left(p_d^{-1}-1\right)H_d(c)\right] + Cov\left(p_d^{-1}-1, H_d(c)\right)\right\}$ is attributed to the variability of sampling probability across strata. The term $[SS_d(c)\times a_d - b_d(c)]^T I_d^{-1}[SS_d(c)\times a_d - b_d(c)]$ corresponds to the reduction in the asymptotic variance attributed to the estimation of sampling probability. Therefore, using estimated sampling weight can lead to improvement in efficiency even if the true sampling probability is known.

Furthermore, based on Theorem 1, we can show the asymptotic variance of the IPW $ROC(t)$ estimator is monotonic decreasing when we add more variables to the model of sampling probabilities in addition to sampling stratum. Therefore, we can further improve asymptotic efficiency of the IPW estimator by adding auxiliary variables as in model (2) after accounting for the information of sampling strata. The proof is provided in supplementary material Appendix B.

In practice, the asymptotic variance $\sigma^2_{ROC(t),IPW}$ can be estimated by substituting each component in the analytical variance with its IPW estimate. For example, $\widehat{SS}_D(c) = \widehat{SEN}(c) = \left[\sum_{i=1}^{N_D}\frac{\delta_{Di}}{\hat{p}_{Di}}I(X_{Di} > c)\right] \Big/ \sum_{i=1}^{N_D}\frac{\delta_{Di}}{\hat{p}_{Di}}$, where $\hat{p}_{D_i}$ is the corresponding estimated sampling probability of case i. Other components can be estimated in a similar manner. Using the asymptotic result in Theorem 1, we can make inference about $ROC(t)$ based on the IPW estimator $\widehat{ROC}^{Aux}_{IPW}(t)$ and its asymptotic variance.

2.1.2 | Estimation of AUC—We next consider the estimation of the area under the ROC curve (AUC) for the biomarker X: $AUC = P(X_D > X_{\bar{D}})$ under a Bernoulli sampling design. The empirical estimator of AUC is

$\widehat{AUC}_{em} = \left[\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}}\delta_{Di}\delta_{\bar{D}j}I\left(X_{Di} > X_{\bar{D}j}\right)\right] \Big/ \left(\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}}\delta_{Di}\delta_{\bar{D}j}\right)$, which can be severely biased in the presence of biased sampling [7]. To correct for the bias of the empirical estimator $\widehat{AUC}_{em}$ under the biased sampling of cases and/or controls, [7] proposed an IPW version estimator

$$\widehat{AUC}_{IPW} = \left[\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}}\frac{\delta_{Di}}{\hat{p}_{Di}}\frac{\delta_{\bar{D}j}}{\hat{p}_{\bar{D}j}}I\left(X_{Di} > X_{\bar{D}j}\right)\right] \Big/ \left(\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}}\frac{\delta_{Di}}{\hat{p}_{Di}}\frac{\delta_{\bar{D}j}}{\hat{p}_{\bar{D}j}}\right), \quad (3)$$

that adopt $\hat{p}_{Di} = \hat{p}_{Di}^{Str}$ and $\hat{p}_{\bar{D}j} = \hat{p}_{\bar{D}j}^{Str}$, the empirically estimated sampling probabilities of the case i and control j, respectively, conditional on their sampling stratum, which we denote here as $\widehat{AUC}_{IPW}^{Str}$. $\widehat{AUC}_{IPW}^{Str}$ has been proven to be asymptotically unbiased and normally distributed.

Similar to our approach for estimating the points on the ROC curve, we consider further improving the efficiency of the IPW AUC estimator by incorporating additional auxiliary variables into the estimation of sampling weights. First, we consider estimating sampling probabilities for cases and controls based on a GLM model conditional on sampling stratum and auxiliary variables as described in Section 2.1.1, i.e., $\hat{p}_{Di} = \hat{p}_{Di}^{Aux}$ and $\hat{p}_{\bar{D}j} = \hat{p}_{\bar{D}j}^{Aux}$. Entering $\hat{p}_{Di}^{Aux}$ and $\hat{p}_{\bar{D}j}^{Aux}$ into (3) leads to the IPW AUC estimator $\widehat{AUC}_{IPW}^{Aux}$.

The asymptotic normality of the IPW AUC estimator developed in [7] holds in general for estimated sampling probabilities based on the GLM model (2) and applies to $\widehat{AUC}_{IPW}^{Aux}$ here. The corresponding results are presented in Theorem 2.

Theorem 2. *Assume* $0 < p_D, p_{\bar{D}} < 1$ *and* $N_D/N \to \lambda \in (0, 1)$ *as the sample size* $N \to \infty$. *Then as* $N \to \infty$, $\sqrt{N}\left(\widehat{AUC}_{IPW}^{Aux} - AUC\right)$ *converges to a normal random variable with mean* 0 *and variance*

$$\sigma^2_{AUC,IPW} = \frac{1}{\lambda}\Sigma_{D,IPW} + \frac{1}{1-\lambda}\Sigma_{\bar{D},IPW}$$

where

$$\Sigma_{d,IPW} = Var(G_d) + E\left[\left(p_d^{-1} - 1\right)G_d^2\right] - AUC\left\{E\left[\left(p_d^{-1} - 1\right)G_d\right] + Cov\left(p_d^{-1} - 1, G_d\right)\right\}$$
$$-(AUC \times a_d - l_d)^T I_d^{-1}(AUC \times a_d - l_d), \text{ for } d = D, \bar{D}$$

in which $G_D = F_{\bar{D}}(X_D)$, $G_{\bar{D}} = 1 - F_D(X_{\bar{D}})$, $l_d = E[I(X_D > X_{\bar{D}})(1/p_d)(\partial_{p_d}/\partial_{\theta_d})]$, a_d, I_d *are defined in Theorem 1 for* $d = D, \bar{D}$.

Again, we observe that $\sigma^2_{AUC,IPW}$ consists of two parts, $\Sigma_{D,IPW}$ and $\Sigma_{\bar{D},IPW}$, which are variance components due to variability of cases and controls, respectively. Additionally, notice that the expression $\Sigma_{d,IPW}$ here has similar structure as the expression $\sigma^2_{d,IPW}$ presented in Theorem 1; both include components for empirical estimator, components due to variability of sampling probability across strata, and components due to estimation of the sampling probability. The difference between $\Sigma_{d,IPW}$ and $\sigma^2_{d,IPW}$ lies in the components related to the classification accuracy measure, e.g., $SS_d(c) = SEN(c)$ or $SPE(c)$ in $\sigma^2_{d,IPW}$ is replaced with AUC in $\Sigma_{d,IPW}$, etc. Similarly, one can show that using estimated sampling probability instead of true probability can result in reduction of $\sum_{d\in\{D,\bar{D}\}}[\lambda+(1-2\lambda)I(d=\bar{D})]^{-1}(AUC\times a_d-l_d)^T I_d^{-1}(AUC\times a_d-l_d)$ in asymptotic variance of $\sqrt{N}\left(\widehat{AUC}^{Aux}_{IPW}-AUC\right)$. We also show that adding auxiliary variables in addition to sampling stratum in modeling the sampling probability can lead to more efficient AUC estimators (proof is given in supplementary material Appendix B).

2.1.3 | Estimation of Partial AUC—In addition to the points on the ROC curve and the AUC, the partial AUC is also an important measure of diagnostic test accuracy when a restricted region under the ROC curve is of interest. For example, in diagnostic testing, it is essential to obtain high sensitivity, i.e. a high true positive rate, to adequately detect individuals with disease. In cancer screening, high specificity, i.e. a low false positive rate, is considered more important. Here, we consider the area under the ROC curve with the restriction that the false positive rate is within a specific range. Such a partial AUC is defined as $pAUC(t_0,t_1)=\int_{t_0}^{t_1}ROC(t)dt$, where the false positive rates fall into the interval (t_0, t_1). Furthermore, we observe that

$$pAUC(t_0,t_1)=\int_{t_0}^{t_1}ROC(t)dt=\int_{t_0}^{t_1}\left\{1-F_D\left[F^{-1}_{\bar{D}}(1-t)\right]\right\}dt$$
$$=\int_{q_1}^{q_0}[1-F_D(u)]f_{\bar{D}}(u)d(u)=P(X_D>X_{\bar{D}},X_{\bar{D}}\in(q_1,q_0)),$$

Where $q_0=F^{-1}_{\bar{D}}(1-t_0)$ and $q_1=F^{-1}_{\bar{D}}(1-t_1)$. [16] proposed a nonparametric partial AUC estimator

$$\widehat{pAUC}_{em}(t_0,t_1)=\left[\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}}\delta_{Di}\delta_{\bar{D}j}I\left(X_{Di}>X_{\bar{D}j},X_{\bar{D}j}\in(q_1,q_0)\right)\right]\Big/\left(\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}}\delta_{Di}\delta_{\bar{D}j}\right).$$

In some circumstances, the quantiles q_0 and q_1 are known. When they are unknown, the empirical quantile estimates are recommended as substitutes. We note that $\widehat{pAUC}_{em}(0,1)$ is

equivalent to the empirical AUC estimator $\widehat{AUC}_{em}$ and expect that the empirical estimator $\widehat{pAUC}_{em}(t_0, t_1)$ can again be seriously biased under a biased sampling scheme for cases and/or controls. Here, we adopt an idea similar to that used for estimating the points on the ROC curve and the AUC and propose the IPW estimator of the partial AUC with different estimated sampling weights:

$$\widehat{pAUC}_{IPW}(t_0, t_1) = \left[\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}} \frac{\delta_{Di}}{\hat{p}_{Di}}\frac{\delta_{\bar{D}j}}{\hat{p}_{\bar{D}j}} I\left(X_{Di} > X_{\bar{D}j}, X_{\bar{D}j} \in (\hat{q}_1, \hat{q}_0)\right)\right] \Big/ \left(\sum_{i=1}^{N_D}\sum_{j=1}^{N_{\bar{D}}} \frac{\delta_{Di}}{\hat{p}_{Di}}\frac{\delta_{\bar{D}j}}{\hat{p}_{\bar{D}j}}\right). \quad (4)$$

In practice, it is most likely that q_0 and q_1 are unknown. Here, we propose to use the IPW estimators $\hat{q}_0 = \hat{F}^{-1}_{\bar{D}, IPW}(1 - t_0)$ and $\hat{q}_1 = \hat{F}^{-1}_{\bar{D}, IPW}(1 - t_1)$, where $\hat{F}_{\bar{D}, IPW}(x)$ is defined in Section 2.1.1. Again, we can use $\hat{p}^{Str}_{Di} \Big/ \hat{p}^{Str}_{\bar{D}j}$ or $\hat{p}^{Aux}_{Di} \Big/ \hat{p}^{Aux}_{\bar{D}j}$ as the estimated sampling probability for cases/controls in the IPW partial AUC estimator (4) and the IPW CDF estimator $\hat{F}_{\bar{D}, IPW}(x)$. We denote the corresponding partial AUC estimators as $\widehat{pAUC}^{Str}_{IPW}$ and $\widehat{pAUC}^{Aux}_{IPW}$. The asymptotic normality of the proposed partial AUC estimators based on the GLM model for sampling probabilities is presented in Theorem 3.

Theorem 3. *Assume* $0 < p_D, p_{\bar{D}} < 1$ *as the sample size* $N \to \infty$. *Then as* $N \to \infty$, $\sqrt{N}\left[\widehat{pAUC}^{Aux}_{IPW}(t_0, t_1) - pAUC(t_0, t_1)\right]$ *converges to a normal random variable with mean* 0 *and variance*

$$\sigma^2_{pAUC, IPW} = \tilde{\sigma}^2_{pAUC, IPW} + \frac{1}{1-\lambda}\left\{\sum_{k=0}^{1} [1 - F_D(q_k)]^2 \sigma^2_{\bar{D}, IPW}(q_k) + 2[Q(q_1) - Q(q_0) - R(q_0, q_1)]\right\}$$

where

$$\tilde{\sigma}^2_{pAUC, IPW} = \frac{1}{\lambda}\Sigma^*_{D, IPW} + \frac{1}{1-\lambda}\Sigma^*_{\bar{D}, IPW},$$

$$\begin{aligned} Q(q_k) = {} & [1 - F_D(q_k)] \times \{Cov[(p_{\bar{D}}^{-1} - 1) J_{\bar{D}}, H_{\bar{D}}(q_k)] \\ & + Cov[J_{\bar{D}}, H_{\bar{D}}(q_k)] - pAUC(t_0, t_1) \times Cov\left[\left(p_{\bar{D}}^{-1} - 1\right), H_{\bar{D}}(q_k)\right] \\ & - [pAUC(t_0, t_1) \times a_{\bar{D}} - s_{\bar{D}}]^T I^{-1}_{\bar{D}} [(1 - t_k) \times a_{\bar{D}} - b_{\bar{D}}(q_k)]\}, \text{ for } k = 0, 1, \end{aligned}$$

$$R(q_0, q_1) = [1 - F_D(q_0)][1 - F_D(q_1)] \times \{Cov[(p_{\bar{D}}^{-1} - 1)H_{\bar{D}}(q_1), H_{\bar{D}}(q_0)] + Cov[H_{\bar{D}}(q_1), H_{\bar{D}}(q_0)] - (1 - t_1) \times Cov[p_{\bar{D}}^{-1} - 1, H_{\bar{D}}(q_0)] - [(1 - t_0) \times a_{\bar{D}} - b_{\bar{D}}(q_0)]^T I_{\bar{D}}^{-1}[(1 - t_1) \times a_{\bar{D}} - b_{\bar{D}}(q_1)]\},$$

in which $J_{\bar{D}} = P(X_{\bar{D}} < X_D, X_{\bar{D}} \in (q_1, q_0))$, $s_d = E[I(X_D > X_{\bar{D}}, X_{\bar{D}} \in (q_1, q_0))\,(1/p_d)\,(\partial p_d/\partial \theta_d)]$, $\Sigma^*_{d,IPW} = Var(J_d) + E\left[\left(p_d^{-1} - 1\right)J_d^2\right] - pAUC\left\{E\left[\left(p_d^{-1} - 1\right)J_d\right] + Cov\left(p_d^{-1} - 1, J_d\right)\right\}$, $H_d(\cdot)$, $-(pAUC \times a_d - s_d)^T I_d^{-1}(pAUC \times a_d - s_d)$

a_d, $b_d(\cdot)$, I_d *are defined in Theorem 1, for* $d = D, \bar{D}$.

Proof of Theorem 3 is presented in supplementary material Appendix C. Note that when $t_0 = 0$ and $t_1 = 1$, $\widehat{pAUC}^{Aux}_{IPW}(t_0, t_1)$ corresponds to $\widehat{AUC}^{Aux}_{IPW}$, and corresponding $\sigma^2_{pAUC,IPW}$ equals $\sigma^2_{AUC,IPW}$. In the expression of $\sigma^2_{pAUC,IPW}$ above, the term $[1 - F_D(q_k)]^2\sigma^2_{\bar{D},IPW}(q_k)$ is caused by the estimation of q_k, the term $Q(q_k)$ is cause by the joint estimation of q_k and the sampling probability, and the term $R(q_0, q_1)$ is caused by the joint estimation of q_0 and q_1. Therefore, when q_0 and q_1 are known, those terms varnish and thus $\sigma^2_{pAUC,IPW}$ is reduced to $\tilde{\sigma}^2_{pAUC,IPW}$. In addition, comparing the expression $\tilde{\sigma}^2_{pAUC,IPW}$ above with $\sigma^2_{AUC,IPW}$ provided in Theorem 2, we notice that the composition of $\Sigma^*_{d,IPW}$ has similar structure as that of $\Sigma_{d,IPW}$. Both of them include components for empirical estimator, components due to variability of sampling probability across strata, and components due to estimation of the sampling probability. The difference between $\Sigma^*_{d,IPW}$ and $\Sigma_{d,IPW}$ lies in the components related to the classification accuracy measure, i.e., G_d, AUC and l_d in $\Sigma_{d,IPW}$ are replaced by J_d, $pAUC$, s_d in $\Sigma^*_{d,IPW}$ respectively, for $d = D, \bar{D}$.

2.2 | Finite-population Stratified Sampling

In this section, we consider estimating the points on the ROC curve, the AUC, and the partial AUC when the finite-population stratified sampling is used in the second phase. In this type of design, fixed numbers of subjects in each stratum are randomly selected in phase two. Suppose that cases and controls among phase-one samples are stratified into K_D and $K_{\bar{D}}$ strata, with N_{Dk_D} and $N_{\bar{D}k_{\bar{D}}}$ samples in each stratum, respectively, where $k_D = 1,2, \cdots, K_D$ and $k_{\bar{D}} = 1, 2, \cdots, K_{\bar{D}}$. In each stratum fixed numbers of n_{Dk_D} cases and $n_{\bar{D}k_{\bar{D}}}$ controls are randomly sampled, with all subjects in the same stratum having the same sampling probability. An example of the finite-population stratified sampling design was given in [17] named the "balanced sampling", where with equal number of samples were acquired across case/control status and covariate-strata.

Let π_{Dk_D} and $\pi_{\bar{D}k_{\bar{D}}}$ be the sampling probabilities of the cases in the stratum k_D and the controls in the stratum $k_{\bar{D}}$, respectively. The sampling probabilities within each case/control sampling stratum can be empirically estimated as $\hat{\pi}_{Dk_D} = n_{Dk_D} \Big/ N_{Dk_D}$ and

$\hat{\pi}_{\bar{D}k_{\bar{D}}} = n_{\bar{D}k_{\bar{D}}} \Big/ N_{\bar{D}k_{\bar{D}}}$. That is, for a case in stratum k_D, the estimated sampling probability conditional on sampling strata is $\hat{p}_D^{Str} = \hat{\pi}_{Dk_D}$; similarly, for a control in stratum $k_{\bar{D}}$, the estimated sampling probability conditional on sampling strata is $\hat{p}_{\bar{D}}^{Str} = \hat{\pi}_{\bar{D}k_{\bar{D}}}$. Let $\delta_{Dk_D,i}$ and $\delta_{\bar{D}k_{\bar{D}},j}$ be the indicators of being sampled in phase two for case i in stratum k_D and control j in stratum $k_{\bar{D}}$, respectively. The IPWAUC estimator (3) with the estimated sampling probability conditional on sampling strata, i.e., $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$, can be rewritten as

$$\widehat{AUC}_{IPW}^{Str} = \frac{\sum_{k_D=1}^{K_D}\sum_{k_{\bar{D}}=1}^{K_{\bar{D}}}\sum_{i=1}^{N_{Dk_D}}\sum_{j=1}^{N_{\bar{D}k_{\bar{D}}}} \frac{\delta_{Dk_D,i}}{\hat{\pi}_{Dk_D}} \frac{\delta_{\bar{D}k_{\bar{D}},j}}{\hat{\pi}_{\bar{D}k_{\bar{D}}}} I\left(X_{Dk_D,i} > X_{\bar{D}k_{\bar{D}},j}\right)}{\sum_{k_D=1}^{K_D}\sum_{k_{\bar{D}}=1}^{K_{\bar{D}}}\sum_{i=1}^{N_{Dk_D}}\sum_{j=1}^{N_{\bar{D}k_{\bar{D}}}} \frac{\delta_{Dk_D,i}}{\hat{\pi}_{Dk_D}} \frac{\delta_{\bar{D}k_{\bar{D}},j}}{\hat{\pi}_{\bar{D}k_{\bar{D}}}}}, \quad (5)$$

which was studied in [7]. When there are additional auxiliary variables available from phase-one samples, we can again use them to further improve efficiency in the estimation of sampling probabilities and compute $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ as in Section 2.1.1. The corresponding IPW AUC estimator $\widehat{AUC}_{IPW}^{Aux}$ can be generated by replacing $\hat{\pi}_{Dk_D}$ and $\hat{\pi}_{\bar{D}k_{\bar{D}}}$ in (5) with with $\hat{p}_{Dk_D,i}^{Aux}$ and $\hat{p}_{\bar{D}k_{\bar{D}},j}^{Aux}$. The proposed IPW $ROC(t)$ estimators (1) and partial AUC estimators (4) for the Bernoulli sampling design can be similarly constructed here for finite-population stratified sampling.

Consider a Bernoulli sampling design in which cases in stratum k_D and controls in stratum $k_{\bar{D}}$ are independently sampled with probability π_{Dk_D} and $\pi_{\bar{D}k_{\bar{D}}}$, respectively, [7] proved that the asymptotic variance of the IPW AUC estimator $\widehat{AUC}_{IPW}^{Str}$ in the finite-population stratified sampling design is equivalent to that in the Bernoulli sampling design if $\hat{\pi}_{Dk_D} \to \pi_{Dk_D}$ and $\hat{\pi}_{\bar{D}k_{\bar{D}}} \to \pi_{\bar{D}k_{\bar{D}}}$. Similar arguments can be used to show the equivalence of asymptotic variance for $\widehat{AUC}_{IPW}^{Aux}$ between these two sampling designs. For the proposed IPW estimators of $ROC(t)$ and partial AUC, the equivalence in asymptotic variance between these two sampling designs can be similarly derived.

R code for estimating various classification accuracy measures under two-phase sampling and for estimating their asymptotic variances can be found in https://research.fhcrc.org/huang/en/software.html.

3 | SIMULATION STUDY

In this section, we conduct simulation studies to evaluate the performance of the proposed IPW estimators for the points on the ROC curve, the AUC, and the partial AUC. Both Bernoulli sampling and finite-population stratified sampling settings are considered.

Let D be a binary disease outcome and set disease prevalence $P(D = 1) = \lambda = 0.1$. We consider one biomarker X and a vector of variables $Z = (V^*, W)^T$ that includes two auxiliary variables V^* and W. X, V^*,and W are jointly normally distributed among the cases and controls. Among the controls, X, V^* and W are each normally distributed with mean 0 and variance 1. Among the cases, the auxiliary variables V^* and W each follow the normal distribution $N(0.5,1)$ and the biomarker X follows the normal distribution $N(1, 1)$. In other words, the marginal distribution of (X, V^*, W) in the population is a mixture of multivariate normal distributions, with weight λ for corresponding multivariate normal distribution among cases and weight $1 - \lambda$ for corresponding multivariate normal distribution among controls. The correlations between X and V^*, X and W, W and V^* are denoted by ρ_{XV^*}, ρ_{XW} and ρ_{WV^*} respectively. Thus, the distribution of (X, V^*, W) in the population is a mixture multivariate normal distribution, i.e., $(X, V^*, W)^T \sim \lambda MVN(\mu_D, \Sigma) + (1 - \lambda)\ MVN(\mu_{\bar{D}}, \Sigma)$, where $\mu_D = (1, 0.5, 0.5)^T$, $\mu_{\bar{D}} = (0, 0, 0)^T$, and $\Sigma = \begin{pmatrix} 1 & \rho_{XV^*} & \rho_{XW} \\ \rho_{XV^*} & 1 & \rho_{WV^*} \\ \rho_{XW} & \rho_{WV^*} & 1 \end{pmatrix}$. The odds ratio for biomarker X is 2.7. Subjects are stratified into two strata based on the value of V^*. Let V be the discrete stratum variable: $V = 1$ if $V^* < \Phi^{-1}(0.5)$, and $V = 2$ if $V^* \geq \Phi^{-1}(0.5)$, where Φ is the CDF of the standard normal distribution.

In the first phase, $N = 2000$, 5000, or 10000 subjects are randomly selected from the population with the disease indicator D and auxiliary variables V^* and W are measured. In the second phase, subsamples are drawn from the phase-one cohort and the biomarker X is measured. In the Bernoulli sampling design, we randomly sample cases with the sampling probability $p_D = 0.5$. For a control, its sampling probability in the stratum $V = v$ is set to be $p_{\bar{D}} = p_D \times P(V = v, D = 1)/P(V = v, D = 0)$. Therefore, in each stratum, we obtain equal numbers of the cases and controls in the second phase on average. In the finite-population stratified sampling design, $n_D = \lambda p_D N$ cases are randomly sampled without replacement. Then, equal numbers of controls as cases in each stratum are randomly selected without replacement. With the phase-one sample size $N = 2000$, 5000, and 10000, the expected numbers of cases and controls sampled in phase-two for Bernoulli sampling and the exact numbers of cases and controls sampled in finite-population stratified sampling are $n_D = n^{\bar{D}} =$ 100,250, and 500, respectively.

For each set of generated data, we estimate the points on the ROC curve, the AUC, and the partial AUC using the empirical method and our proposed IPW method with two different types of estimated sampling probabilities for cases/controls: $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$ (conditional on V only), and $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ (based on a linear logistic regression of sampling probability conditional on both V and Z). To evaluate the performance of these estimation methods, we

examine the averaged estimate, bias, sample variance, the median of estimated analytical variance and coverage of 95% confidence interval based on 5000 Monte-Carlo simulations. Let $\hat{y}_k$ be an estimator of the corresponding measure in the k-th replication and y be the true value. The averaged estimate, bias, and sample variance are defined as $\bar{y} = \frac{1}{r}\sum_{k=1}^{r}\hat{y}_k$, $\bar{y} - y$, and $\frac{1}{r-1}\sum_{k=1}^{r}(\hat{y}_k - \bar{y})^2$, respectively. We also estimate variance based on our developed analytical variance formulas and compute its median across Monte-Carlo simulations. Furthermore, 1 the estimated analytical variances are adopted to construct the 95% Wald confidence interval of each estimate assuming approximate normality of the estimator after logit-transformation. We report the percent of times such confidence intervals contain the true value.

The simulation results with $\rho_{XV^*} = 0.5$, $\rho_{XW} = 0.5$, and $\rho_{WV^*} = 0.1$ are presented in Tables 1 -2 for finite-population stratified sampling and in supplementary material Appendix D Tables 1-2 for Bernoulli sampling design. We can see that for both sampling designs, the IPW estimators outperform the empirical estimators. The empirical estimators are generally biased and the corresponding CIs have lower coverage compared to the nominal level. In contrast, the IPW methods are asymptotically unbiased and the coverage of the 95% CIs is very close to the nominal level. We also observe that medians of the estimated analytical variances of each classification accuracy measure are very close to their sample variances. Moreover, when comparing across the two IPW methods with different types of estimated weights, the IPW estimators for $ROC(t)$, AUC, and $pAUC(t_0, t_1)$ with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ have smaller variances than the corresponding estimators with $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$. These findings demonstrate that efficiency in estimating biomarker classification accuracy is improved when auxiliary variables in phase one are utilized in estimating the sampling probability through the generalized linear model.

To further compare the efficiencies of the two IPW estimators with different types of estimated weights, we compute the efficiency of the IPW estimator with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ relative to the estimator with $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$, for scenarios with varying degrees of correlation between X and W: $\rho_{XW} = 0.1$, 0.3, 0.5 and 0.7. Here, the efficiency of one estimator relative to another is defined as the ratio of the variance of the latter relative to the variance of the former. The simulation results with $N = 5000$, $\rho_{XV^*} = 0.3$, $\rho_{WV^*} = 0$ for both Bernoulli sampling and finite-population stratified sampling are shown in Table 3. For each value of ρ_{XW} in {0.1, 0.3, 0.5, 0.7}, we observe higher efficiency in the IPW estimator with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ compared to the one with $\hat{p}_D^{Str} \Big/ \hat{p}_{\bar{D}}^{Str}$. Furthermore, higher correlation between the biomarker X and the auxiliary variable W is associated with a higher efficiency of the IPW estimator with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$. For example, in finite-population stratified sampling, the efficiency gain in estimating AUC by using $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ improves from 3.6% for $\rho_{XW} = 0.1$ to 49.5% for $\rho_{XW} = 0.7$. For various points on the ROC curve and for pAUC, we can also see appreciable efficiency gains when ρ_{XW} is not very small. This suggests that considerable improvements

in the precision of classification performance estimators could be achieved via the use of auxiliary variables in estimating sampling weights when the auxiliary variables have good correlations with the biomarker. Note that in our simulation settings, even when X and W have minimal correlation there is no loss of efficiency when incorporating auxiliary variables in estimating sampling weights, although the efficiency gain can be minor in such scenarios.

Finally, to investigate how estimation efficiency gain with auxiliary variables changes with the classification performance of the biomarker X, we repeat the above simulation with varying means of X among cases, i.e $\mu_{X_D} = 0$, 0.6 or 1.5, corresponding to an odds ratio of 1, 1.8, and 4.5, and an AUC of 0.5, 0.66, and 0.86 respectively. We compare the efficiencies across the three IPW methods. The simulation results are summarized in supplementary material Appendix D Table 3. We observe a very similar pattern of efficiency gain with the use of auxiliary variable as the scenario with $\mu_{X_D} = 1$. Thus, the efficiency gain is quite robust to classification performance of X.

4 | EXAMPLE

4.1 | Renal Artery Stenosis Study Example

We now apply the proposed estimators for classification accuracy to evaluate serum creatinine as a classification biomarker for renal artery stenosis in patients with therapy-resistant hypertension, using data from a prospective renal artery stenosis study conducted at 26 departments of internal medicine throughout the Netherlands [18]. The original study by [18] included 477 patients with hypertension. In our analysis, we include 426 patients with complete covariate information, including age (Z_1), sex (Z_2), smoking history (Z_3), body mass index (Z_4), recent onset of hypertension (Z_5), abdominal bruit (Z_6), atherosclerotic vascular disease indicator (Z_7), and hypercholesterolemia (Z_8). The prevalence of renal artery stenosis (D) was 23.0%. The biomarker of interest, log-transformed serum creatinine concentration (X), was also measured in this cohort.

To demonstrate the application of our proposed methods, we conduct a finite-population stratified case-control sampling from the study cohort based on three age strata generated by the first three quantiles of age distribution among controls, i.e., ≤ 45,46 - 56, ≥ 57 years. In the phase-one study cohort, among each age stratum, there are 16, 28, and 54 cases and 112,112, and 104 controls respectively. The phase-two case-control sample includes all 98 cases in the cohort. Equal numbers of controls as cases within each stratum are randomly sampled. To estimate the points on the ROC curve, the AUC, and the partial AUC of creatinine, we compute the empirical estimators and our proposed IPW estimators with the two choices of estimated sampling weights described in Section 2. For the IPW method with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$, we include Z_1, Z_2, Z_3, Z_4, Z_5, Z_6, Z_7, Z_8 and $Z_2 \times Z_3$ as the auxiliary variables, all of which are considered to be risk factors for hypertension [18]. Among these risk factors, their correlation with log(creatinine) ranges from −0.001 (recent onset of hypertension) to 0.41 (sex).

We summarize the classification performance estimates, estimated variances based on the analytical formula, 95% confidence intervals, and their lengths in Table 4. We observe an obvious difference in the estimates between the empirical and IPW methods, suggesting the bias of the empirical estimates. The empirical estimates based on this stratified case-control sample tend to under-estimate classification performance of creatinine. Furthermore, when comparing across the IPW methods with two different types of estimated sampling weights, the results demonstrate an apparent improvement in estimation precision when we apply estimated weights incorporating auxiliary variable information. In this example, the IPW estimator with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$ shows the best efficiency regarding the estimation of $ROC(t)$, AUC and $pAUC(t_0, t_1)$.

4.2 | Prostate Cancer Study Example

In the second data example, we apply our methodology to assess a urine biomarker, Prostate Cancer Antigen 3 (PCA3), for early detection of prostate cancer. The study cohort comes from a prospective study conducted by the Early Detection Research Network. The original cohort involved 570 men, with a prostate cancer (D) prevalence of 36.6% [19]. Here, we include 532 men with complete covariate information, including patient age, suspicious digital rectal exam (DRE), prostate gland volume, serum prostate-specific agent (PSA), family history of prostate cancer and whether or not the patient had a prior negative prostate biopsy as our phase-one study cohort.

To illustrate the application of our methodology, we adopt a finite-population stratified case-control sampling design from the phase-one cohort based on three age strata generated by the first three quantiles of age distribution among controls. In the phase-one cohort, among each age stratum, there are 48, 54, and 84 cases and 116, 115, and 115 controls, respectively. In the second phase, 120 cases are randomly sampled, resulting in 30, 38, and 52 cases in each age stratum. Then equal numbers of controls as cases within each age stratum are randomly drawn. We estimate the points on the ROC curve, AUC, and partial AUC based on PCA3 using the empirical method and our proposed IPW methods with two different types of estimated sampling weights. For the IPW method with $\hat{p}_D^{Aux} \Big/ \hat{p}_{\bar{D}}^{Aux}$, we include all covariates mentioned above as the auxiliary variables. Among those variables, the highest correlation is observed between age and log(PCA3) (0.38), while other variables have correlation with log(PCA3) smaller than 0.1.

The results are reported in supplementary material Appendix E Table 4. In this example, the empirical estimators again appear to under-estimate the biomarker's performance compared to the IPW methods. The two IPW estimators, on the other hand, are close to each other. A small efficiency gain is observed when the estimated weights using auxiliary variables are applied to the IPW estimators.

5 | CONCLUDING REMARKS

In this paper, we developed an IPW-based method for evaluating a biomarker's classification accuracy with respect to the points on the ROC curve, the AUC, and the partial AUC in two-phase case-control sampling designs. We investigated two-phase sampling designs where the

phase-one sample is a simple random sample from the target population and in the second phase cases and control are sampled with probabilities depending on covariate strata [20]. When cases and controls in the second phase are not simply randomly sampled from their corresponding populations, we showed that traditional empirical classification accuracy estimators can be seriously biased, leading to invalid inference regarding the biomarker's performance. While the importance of accounting for biased sampling in biomarker evaluation is well recognized, systematic research regarding how to best adjust for biased sampling for various classification performance measures and their asymptotic properties has been lacking. Recently, the idea of inverse probability weighting was adopted to develop an unbiased estimator of AUC, which utilizes sampling probability estimates conditional on the sampling strata alone [7]. While being an extension of the work described by [7], the current paper covers much broader ground. In particular, we proposed IPW estimators for characterizing a biomarker's performance over a full list of commonly used classification measures, including the points on the ROC curve, the AUC, and the partial AUC. More importantly, we further improved the efficiency of the IPW estimator by estimating sampling weights via the use of auxiliary variables available in phase one in addition to the use of sampling stratum. The more complicated model we adopt for modeling sampling probability includes sampling strata as part of the covariates set. As a result, the true sampling probability model that depends on sampling strata only is nested within the more complicated model. Both models lead to consistent estimates of sampling weights, which are required for valid performance of the IPW estimators. We developed analytical variance formulas for the proposed IPW-based estimators that are applicable to both Bernoulli sampling designs and finite-population stratified sampling designs, which are useful for making inference about classification accuracy. In addition, these estimators can be valuable in guiding biomarker study design, e.g., by suggesting auxiliary covariates to collect in the first phase for potential help with estimation efficiency. Through extensive numerical studies, we showed that appreciable efficiency gain can be achieved by using auxiliary variables in modeling sampling probability in a generalized linear model, especially when there is a strong correlation between the auxiliary variables and the biomarker of interest. These results are currently lacking in applied biomarker research and we anticipate this paper could serve as a useful reference and guideline for improving the practice of biomarker evaluation.

In practice, we recommend incorporating auxiliary variables to estimate weights for the IPW estimators when there exist easy-to-collect auxiliary variables with some correlation with the biomarker of interest. On the one hand, appreciable efficiency gain can be achieved when correlation between the biomarker and auxiliary variable is not too small (e.g. a correlation of a level ~ 0.3); on the other hand, even when correlation is minimal, incorporating auxiliary variables would not hurt efficiency and might still lead to some minor efficiency gain in finite samples. In practice, it might happen that a few irrelevant auxiliary variables are also included when estimating the sampling weights. For practical sample size like 100 cases and 100 controls, the sampling weight estimated including these auxiliary variables can still lead to better efficiency compared to using the discrete sampling strata alone for weights estimation, as illustrated by the numerical results presented in Supplementary Appendix F. We found it particularly desirable to estimate weight by modeling the sampling

probabilities for cases/controls because of the simplicity in implementing the procedure using standard statistical software and the fact that the asymptotic variance of the IPW estimator is monotonic decreasing with the addition of more variables. Validity of the IPW estimator relies on correct specification of the model for sampling weight. In the problem settings we consider in this paper, the true sampling strata are known and our sampling model including the sampling strata in addition to auxiliary variables is guaranteed to be correctly specified. In general, if the model for sampling weight is misspecified, the resulting classification performance estimator can be biased with undercoverage problem in corresponding confidence interval, as demonstrated in numerical studies presented in Supplementary Appendix G. Interestingly, where sampling strata are derived from a continuous auxiliary variable, including the continuous auxiliary variable alone but not the sampling strata when estimating sampling weight might still lead to biased estimate in some settings. This highlights the importance of having correctly specified sampling model, which is achieved in our proposed estimator by always including the known sampling strata as a part of the sampling weight model in addition to auxiliary variables.

The current paper focuses on the two-phase sampling design where a simple random sample representing the target population is obtained in the first phase for measuring disease status and easily-collected covariates, a type of design frequently seen in biomarker research, with the IPW method taking into account biased sampling of biomarker in the second phase. The proposed IPW estimator for classification performance would also work in two-phase designs where cases and controls are randomly sampled from their corresponding populations in the first phase, as considered in [21], given the fact that classification performance is defined based on the comparison between case and control distributions. [22] considered the estimation of sensitivity and specificity under a two-phase sampling design for a different problem setting of verification bias correction. In their problems, a simple random sample or stratified sample is performed in the first phase to measure a standard test, in the second-phase all test-positive and a fraction of test-negative individuals have their disease status verified as well as the biomarker of interest measured. They assumed known population proportion of diseased/non-diseased individuals in each stratum and the proportion of sampled test-negative individuals, and proposed IPW estimators based on the known weights. While our current paper addressed the biased sampling design also using IPW-type estimators, we focused on using estimated weights to achieve a better estimation efficiency. It can be seen from the theoretical results that asymptotic variances of our proposed IPW estimators for various classification accuracy measures can be reduced by using estimated weights and that auxiliary variables not affecting the true sampling probability can be included in the modeling to further improve efficiency.

Finally, it is worth mentioning that the IPW methods we developed in this paper have general applications in biomarker research. While our current work focuses on estimation of various classification measures of a single marker and the usefulness of auxiliary variables in estimating the weights, the same IPW weights can be adopted to make inference about the comparison between two biomarkers with respect to the points on the ROC curve and the (partial) AUC. More generally, the asymptotic normality of the IPW estimators of marker performance allows testing the equivalence of diagnostic accuracy among multiple biomarkers using e.g. a Hotelling’s t-squared statistics [23]. Our paper focuses on estimation

of marker distributions among cases and controls nonparametrically in constructing an ROC curve, but the IPW weights developed in this paper can be similarly applied when a smooth ROC curve estimator is desired in two-phase sampling designs. One way to derive a smoothed ROC curve is to assume a parametric model on the marker distributions among cases and controls, and the IPW weights can be applied to cases and controls separately for estimating corresponding parametric marker distributions. IPW weights can also be applied to account for biased sampling in semi-parametric ROC modeling approaches that assume a parametric form on the ROC curve itself but not on the marker distributions. For example, recognizing the equivalence between the ROC curve and the cumulative distribution function of a case placement value $U_D = 1 - F_{\bar{D}}(X_D)$, i.e. $P(U_D \leq u) = ROC(u)$, [24] proposed a pseudo-likelihood procedure that estimates the placement value for each case $\hat{U}_D = 1 - \hat{F}_{\bar{D}}(X_D)$ first, and then estimates the ROC curve by maximizing a pseudo-likelihood of observed placement values. In two-phase sampling design, this can be extended by first applying IPW weights for controls in estimating $F_{\bar{D}}$ and subsequently placement values for all cases, and then maximizing an inverse-probability weighted pseudo-likelihood based on case placement value estimates.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGMENTS

This work was supported by the U.S. National Institutes of Health under award R01 GM106177-01.

References

[1]. Pepe MS. The statistical evaluation of medical tests for classification and prediction. Medicine; 2003.

[2]. Zhou XH, McClish DK, Obuchowski NA. Statistical methods in diagnostic medicine. John Wiley & Sons; 2009.

[3]. McClish DK. Analyzing a portion of the ROC curve. Medical Decision Making. 1989;9(3):190–195. [PubMed: 2668680]

[4]. Thompson ML, Zucchini W. On the statistical analysis of ROC curves. Statistics in medicine. 1989;8(10):1277–1290. [PubMed: 2814075]

[5]. Pepe MS, Feng Z, Janes H, Bossuyt PM, Potter JD. 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]

[6]. Pepe MS, Fan J, Seymour CW, Li C, Huang Y, Feng Z. Biases introduced by choosing controls to match risk factors of cases in biomarker research. Clinical chemistry. 2012;58(8):1242–1251. [PubMed: 22730452]

[7]. Huang Y. Evaluating and comparing biomarkers with respect to the area under the receiver operating characteristics curve in two-phase case-control studies. Biostatistics. 2016;17(3):499–522. [PubMed: 26883772]

[8]. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. Journal of the American statistical Association. 1994;89(427):846–866.

[9]. Breslow NE, Wellner JA. Weighted Likelihood for Semiparametric Models and Two-phase Stratified Samples, with Application to Cox Regression. Scandinavian Journal of Statistics. 2007;34(1):86–102.

[10]. Saegusa T, Wellner JA. Weighted likelihood estimation under two-phase sampling. Annals of statistics. 2013;41(1):269. [PubMed: 24563559]

[11]. Manski CF, Lerman SR. The estimation of choice probabilities from choice based samples. Econometrica: Journal of the Econometric Society. 1977;:1977–1988.

[12]. Neyman J Contribution to the theory of sampling human populations. Journal of the American Statistical Association. 1938;33(201):101–116.

[13]. Wieand S, Gail MH, James BR, James KL. Afamily of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. Biometrika. 1989;76(3):585–592.

[14]. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. Journal of the American statistical Association. 1952;47(260):663–685.

[15]. Hsieh F, Turnbull BW. Nonparametric and semiparametric estimation of the receiver operating characteristic curve. The annals of statistics. 1996;:25–40.

[16]. Dodd LE, Pepe MS. Partial AUC estimation and regression. Biometrics. 2003;59(3):614–623. [PubMed: 14601762]

[17]. Breslow NE, Cain KC. Logistic regression for two-stage case-control data. Biometrika. 1988;75(1):11–20.

[18]. Krijnen P, Jaarsveld BC, Steyerberg EW, Schalekamp MA, Habbema JDF, others. A clinical prediction rule for renal artery stenosis. Annals of Internal Medicine. 1998;129(9):705–711. [PubMed: 9841602]

[19]. Deras IL, Aubin SMJ, Blase A, et al. PCA3: a molecular urine assay for predicting prostate biopsy outcome. The Journal of urology. 2008;179(4):1587–1592. [PubMed: 18295257]

[20]. Scott AJ, Wild CJ. Fitting regression models to case-control data by maximum likelihood. Biometrika. 1997;84(1):57–71.

[21]. Breslow NE, Holubkov R. Maximum Likelihood Estimation of Logistic Regression Parameters under Two-phase, Outcome-dependent Sampling. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 1997;59(2):447–461.

[22]. Obuchowski NA, Zhou XH. Prospective studies of diagnostic test accuracy when disease prevalence is low. Biostatistics. 2002;3(4):477–492. [PubMed: 12933593]

[23]. Hotelling H The generalization of Student’s ratio. In: Springer 1992 (pp. 54–65).

[24]. Pepe MS, Cai T. The analysis of placement values for evaluating discriminatory measures. Biometrics. 2004;60(2):528–535. [PubMed: 15180681]

TABLE 1

Estimate, bias, variance, median of estimated variance ($Med(\widehat{Var})$), and coverage of 95% confidence interval (CI) of $ROC(t)$ estimator using the empirical method, IPW method with the estimated sampling probabilities $\hat{p}^{Str}$ and $\hat{p}^{Aux}$, for scenarios where $\rho_{XV^*} = 0.5$, $\rho_{XW} = 0.5$, $\rho_{WV^*} = 0.1$, $\mu_{X_{\bar{D}}} = 0$, and $\mu_{X_D} = 1$ in **finite-population stratified sampling**.

Method	$n_D = n_{\bar{D}}$	Estimate	Bias × 100	Var × N	$Med(\widehat{Var}) \times N$	Coverage of 95% CI
			$ROC(0.1) = 0.3891$			
$\widehat{ROC}_{em}$	100	0.3406	−4.8593	11.046	11.766	94.2%
	250	0.3390	−5.0140	11.312	11.856	85.2%
	500	0.3379	−5.1204	11.449	12.000	71.8%
$\widehat{ROC}^{Str}_{IPW}$	100	0.3894	0.0270	11.918	11.236	95.3%
	250	0.3881	−0.1045	11.862	11.750	95.6%
	500	0.3878	−0.1364	12.223	11.848	94.8%
$\widehat{ROC}^{Aux}_{IPW}$	100	0.3885	−0.0606	10.775	9.834	94.5%
	250	0.3883	−0.0857	10.540	10.291	95.4%
	500	0.3876	−0.1524	10.666	10.390	94.9%
			$ROC(0.2) = 0.5629$			
$\widehat{ROC}_{em}$	100	0.5058	−5.7108	10.247	11.178	90.0%
	250	0.5052	−5.7711	10.244	11.267	78.1%
	500	0.5042	−5.8691	10.484	11.243	59.6%
$\widehat{ROC}^{Str}_{IPW}$	100	0.5613	−0.1617	10.654	10.274	95.5%
	250	0.5612	−0.1676	10.402	10.469	95.6%
	500	0.5606	−0.2350	10.773	10.530	94.9%
$\widehat{ROC}^{Aux}_{IPW}$	100	0.5607	−0.2252	9.123	8.771	95.0%
	250	0.5614	−0.1501	9.009	8.946	95.5%
	500	0.5604	−0.2512	9.193	8.997	94.5%
			$ROC(0.5) = 0.8413$			
$\widehat{ROC}_{em}$	100	0.7990	−4.2309	5.166	5.691	86.0%
	250	0.7988	−4.2505	5.286	5.688	73.0%
	500	0.7984	−4.2945	5.271	5.721	52.2%
$\widehat{ROC}^{Str}_{IPW}$	100	0.8387	−0.2684	4.749	4.520	96.0%
	250	0.8396	−0.1737	4.690	4.594	95.2%
	500	0.8399	−0.1401	4.632	4.620	94.6%
$\widehat{ROC}^{Aux}_{IPW}$	100	0.8385	−0.2860	4.170	3.967	95.9%
	250	0.8398	−0.1578	4.251	4.016	94.6%
	500	0.8399	−0.1489	4.078	4.030	95.0%

TABLE 2

Estimate, bias, variance, median of estimated variance ($Med(\widehat{Var})$) and coverage of 95% confidence interval (CI) of AUC and $pAUC(t_0, t_1)$ estimators using the empirical method, IPW method with the estimated sampling probabilities $\hat{p}^{Str}$ and $\hat{p}^{Aux}$, for scenarios where $\rho_{XV^*} = 0.5$, $\rho_{XW} = 0.5$, $\rho_{WV^*} = 0.1$, $\mu_{X\bar{D}} = 0$, and $\mu_{XD} = 1$ in **finite-population stratified sampling**.

Method	$n_D = n_{\bar{D}}$	Estimate	Bias × 100	Var × N	$Med(\widehat{Var}) \times N$	Coverage of 95% CI
			$AUC = 0.7602$			
$\widehat{AUC}_{em}$	100	0.7258	−3.4428	2.1767	2.4639	83.7%
	250	0.7265	−3.3782	2.1105	2.4791	65.8%
	500	0.7264	−3.3857	2.2695	2.4807	38.7%
$\widehat{AUC}^{Str}_{IPW}$	100	0.7597	−0.0564	2.0841	2.1019	95.1%
	250	0.7604	0.0105	2.0643	2.0993	95.5%
	500	0.7603	0.0057	2.2095	2.1025	94.3%
$\widehat{AUC}^{Aux}_{IPW}$	100	0.7600	−0.0244	1.6382	1.6100	95.0%
	250	0.7605	0.0223	1.6015	1.6149	94.9%
	500	0.7603	0.0014	1.7089	1.6146	94.5%
			$pAUC(0,0.1) = 0.0244$			
$\widehat{pAUC}_{em}$	100	0.0198	−0.4535	0.0632	0.0685	95.0%
	250	0.0203	−0.4043	0.0663	0.0695	87.8%
	500	0.0204	−0.4002	0.0654	0.0701	75.5%
$\widehat{pAUC}^{Str}_{IPW}$	100	0.0230	−0.1351	0.0688	0.0642	94.9%
	250	0.0236	−0.0721	0.0716	0.0681	94.8%
	500	0.0238	−0.0566	0.0711	0.0695	94.7%
$\widehat{pAUC}^{Aux}_{IPW}$	100	0.0229	−0.1464	0.0629	0.0613	95.4%
	250	0.0236	−0.0713	0.0630	0.0641	95.5%
	500	0.0238	−0.0591	0.0618	0.0656	95.5%
			$pAUC(0,0.2) = 0.0726$			
$\widehat{pAUC}_{em}$	100	0.0619	−1.0735	0.2692	0.2975	92.6%
	250	0.0625	−1.0053	0.2787	0.2990	81.2%
	500	0.0626	−0.9949	0.2785	0.3016	62.8%
$\widehat{pAUC}^{Str}_{IPW}$	100	0.0698	−0.2757	0.2874	0.2728	94.5%
	250	0.0711	−0.1515	0.2899	0.2809	95.0%
	500	0.0714	−0.1212	0.2965	0.2841	94.6%
$\widehat{pAUC}^{Aux}_{IPW}$	100	0.0696	−0.2949	0.2489	0.2544	95.9%
	250	0.0711	−0.1496	0.2478	0.2615	95.8%
	500	0.0713	−0.1253	0.2459	0.2634	95.5%

TABLE 3

Efficiency comparison of the IPW $ROC(t)$, AUC and $pAUC(t_0, t_1)$ estimators with two different types of estimated sampling weights, for scenarios where $n_D = n_{\bar{D}} = 500$, $\rho_{XV^*} = 0.3$, $\rho_{WV^*} = 0$, $\mu_{X\bar{D}} = 0$, and $\mu_{X_D} = 1$.

	Parameter	True performance	Bernoulli Sampling ρ_{XW}				Finite-population stratified sampling ρ_{XW}			
			0.1	0.3	0.5	0.7	0.1	0.3	0.5	0.7
$\hat{p}^{Aux}_{\text{VS}}$. $\hat{p}^{Str}$	$ROC(0.1)$	0.3891	1.003	1.037	1.103	1.192	1.008	1.034	1.105	1.186
	$ROC(0.2)$	0.5629	1.018	1.065	1.132	1.261	1.018	1.067	1.133	1.268
	AUC	0.7602	1.033	1.102	1.231	1.477	1.036	1.099	1.223	1.496
	$pAUC(0, 0.1)$	0.0244	1.005	1.042	1.106	1.175	1.021	1.029	1.113	1.183
	$pAUC(0, 0.2)$	0.0726	1.027	1.063	1.138	1.275	1.034	1.044	1.137	1.270

TABLE 4

Renal artery stenosis study example: Estimate, variance, 95% confidence interval (CI) and corresponding length of 95% CI for $ROC(t)$, AUC and $pAUC(t_0, t_1)$ estimators using the empirical method and IPW methods with the estimated sampling probabilities $\hat{p}^{Str}$ and $\hat{p}^{Aux}$.

	Method	Est	Var×N	95% CI	Length of 95% CI
ROC (0.1)	Empirical	0.296	4.632	(0.092, 0.500)	0.409
	IPW with $\hat{p}^{Str}$	0.367	3.336	(0.194, 0.541)	0.347
	IPW with $\hat{p}^{Aux}$	0.388	2.081	(0.251, 0.525)	0.274
ROC (0.2)	Empirical	0.449	2.263	(0.306, 0.592)	0.286
	IPW with $\hat{p}^{Str}$	0.480	2.189	(0.339, 0.620)	0.281
	IPW with $\hat{p}^{Aux}$	0.500	1.805	(0.372, 0.628)	0.255
AUC	Empirical	0.665	0.619	(0.590, 0.740)	0.149
	IPW with $\hat{p}^{Str}$	0.682	0.677	(0.604, 0.760)	0.156
	IPW with $\hat{p}^{Aux}$	0.709	0.565	(0.638, 0.781)	0.143
$pAUC$(0, 0.1)	Empirical	0.014	0.012	(0.003, 0.024)	0.021
	IPW with $\hat{p}^{Str}$	0.017	0.028	(0.001, 0.033)	0.032
	IPW with $\hat{p}^{Aux}$	0.018	0.024	(0.004, 0.033)	0.029
$pAUC$(0, 0.2)	Empirical	0.053	0.063	(0.029, 0.077)	0.048
	IPW with $\hat{p}^{Str}$	0.068	0.070	(0.043, 0.093)	0.050
	IPW with $\hat{p}^{Aux}$	0.069	0.057	(0.047, 0.092)	0.045