Published in final edited form as:

Biometrics. 2013 March ; 69(1): 91–100. doi:10.1111/biom.12001.

Covariate adjustment in estimating the area under ROC curve with partially missing gold standard

Danping Liu*

National Alzheimer's Coordinating Center, University of Washington, Seattle, WA 98195

Xiao-Hua Zhou*

Northwest HSR&D Center of Excellence, VA Puget Sound Health Care System, Seattle, WA 98108

Department of Biostatistics, University of Washington, Seattle, WA 98195

Summary

In ROC analysis, covariate adjustment is advocated when the covariates impact the magnitude or accuracy of the test under study. Meanwhile, for many large scale screening tests, the true condition status may be subject to missingness because it is expensive and/or invasive to ascertain the disease status. The complete-case analysis may end up with a biased inference, also known as “verification bias”. To address the issue of covariate adjustment with verification bias in ROC analysis, we propose several estimators for the area under the covariate-specific and covariate-adjusted ROC curves (AUC_X and $AAUC$). The AUC_X is directly modelled in the form of binary regression, and the estimating equations are based on the U statistics. The $AAUC$ is estimated from the weighted average of AUC_X over the covariate distribution of the diseased subjects. We employ reweighting and imputation techniques to overcome the verification bias problem. Our proposed estimators are initially derived assuming that the true disease status is missing at random (MAR), and then with some modification, the estimators can be extended to the not-missing-at-random (NMAR) situation. The asymptotic distributions are derived for the proposed estimators. The finite sample performance is evaluated by a series of simulation studies. Our method is applied to a data set in Alzheimer's disease research.

Keywords

Alzheimer's disease; area under ROC curve; covariate adjustment; U statistics; verification bias; weighted estimating equations

1. Introduction

In ROC analysis, covariate adjustment is advocated when the magnitude and/or accuracy of the test result depends on patient characteristics. This is analogous to adjusting for confounders and effect modifiers in linear regression. In this paper, we focus on the covariate-adjustment in estimating the area under ROC curve (AUC), which measures the probability that a diseased subject (case) gets a larger test score than a healthy subject (control). Covariate-specific AUC (AUC_X) is commonly used in the literature for examining

*dpliu@uw.edu

*azhou@uw.edu

7. Supplementary Materials Supplementary Web Appendices referenced in Sections 2–5 are available with this paper at the Biometrics website on Wiley Online Library.

the diagnostic accuracy within a subpopulation stratified by the covariates. Janes and Pepe (2009) proposed a new covariate-adjusted ROC curve and AUC (AROC and AAUC), which are weighted averages of the covariate-specific ROC curve (ROC_x) and AUC_x , respectively. AAUC conveniently summarizes a test's overall accuracy by a single number, while taking the covariates information into consideration.

For large-scale observational studies, the gold standard for a patient's true disease status may not be available, due to high cost or harm to the patient. The decision of disease verification could be made by physicians or the patients themselves, and is often associated with the test results and other patient characteristics. The estimated ROC curve and AUC only using verified subjects can be biased, known as "verification bias" (Begg and Greenes, 1983). Since the non-verified subjects can be regarded as missing disease status, the missing data framework is introduced to adjust for the verification bias. The verification process is said to be missing at random (MAR) if the probability of disease verification is only affected by the observed variables, and not missing at random (NMAR) if the verification is associated with the missing disease status itself conditioning on all the observed variables.

Recently, much attention has been paid to correct verification bias for the continuous test. Under the MAR assumption, Alonzo and Pepe (2005) first proposed several empirical estimators for the ROC curve. He *et al.* (2009) further derived the AUC estimator using the inverse probability weighting approach. Rotnitzky *et al.* (2006) considered an NMAR assumption and proposed a doubly robust AUC estimator, and Fluss *et al.* (2009) later developed the ROC curve estimator under a similar framework. However, both papers had to assume that a "nonignorable parameter" (log odds ratio of verification for diseased versus healthy subjects) was known and performed a sensitivity analysis. Alternatively, Liu and Zhou (2010) proposed to estimate the nonignorable parameter from the data using the score equations, and then construct several empirical ROC curve and AUC estimators. The biggest limitation of the above five recent papers is that they all performed the ROC analysis on the whole population, and none of them could adjust for the covariates effect on the classification accuracy. Page and Rotnitzky (2010), and Liu and Zhou (2011) are the only two papers, which considered bias-corrected covariate-specific ROC (ROC_x) curve estimator for continuous test: the former proposed a fully parametric model and the latter used a semiparametric framework. To our knowledge, the AUC_x and AAUC estimators under verification bias have not been studied yet.

The main contributions of this paper are: (1) we propose the U-statistic type estimating equations for verification-bias corrected AUC_x and AAUC; (2) we prove the asymptotic theories for the new estimators. Theoretically, once we have the estimated ROC_x curve, AUC_x could be computed. For example, using the ROC_x estimator in Liu and Zhou (2011), one may integrate the ROC curve over $[0, 1]$. However, as the link and baseline function of the ROC_x estimator are both nonparametric functions, AUC_x may not have an explicit expression. In addition, the covariate effects in both Liu and Zhou (2011) and Page and Rotnitzky (2010) are interpreted as the effect on the mean test result. But in many situations, one may wish to find out whether and how the diagnostic accuracy itself is affected by the covariates, and hence it is more relevant to model AUC_x directly. The idea of our approach is, the regression model assumption on AUC_x is made for the full data, and then several reweighting techniques are used to correct for the verification bias. The reweighting methods are first derived under the MAR assumption and then extended to the NMAR situation. Subsequently, the AAUC estimators can be derived as a weighted average of AUC_x . The weights depend on the covariate distribution for diseased subjects, and are estimated empirically. The asymptotic distributions for the AUC_x and AAUC estimators are derived from the U statistics theory.

The paper is organized as follows. In Section 2, we propose the verification bias-corrected estimators of AUC_x . In Section 3, we construct a weighted average of the estimated AUC_x as an estimator of AAUC. Several simulation studies are presented in Section 4, followed by a real example from Alzheimer's disease research in Section 5. Finally we make the concluding remarks in Section 6.

2. Estimation for Covariate-Specific AUC (AUC_x)

Let T_i , D_i , V_i and X_i denote the continuous test result, binary disease status ($D_i = 1$ if diseased and 0 if healthy), binary verification status ($V_i = 1$ if D_i is observed and 0 if D_i is missing), and patient-level characteristics for subject i . Without loss of generality, we assume that a greater value of T_i is more indicative of disease. The subscript i is omitted sometimes if there is no confusion. In this section we first discuss the model setting and assumptions, then we propose to use the weighted estimating equations to correct for the verification bias and obtain the estimated AUC_x , and we present the asymptotic results in the end.

2.1 Model Assumption

The covariate-specific AUC measures the diagnostic accuracy of a test in a subpopulation defined by covariate value $\mathbf{x}$. Similar to the ordinary unadjusted AUC, AUC_x is interpreted as the probability that a case has a greater test result than a control when they share the common covariate $\mathbf{x}$, i.e.,

$$\nu(\mathbf{x}) = \Pr(T_2 > T_1 | D_2 = 1, D_1 = 0, \mathbf{X}_2 = \mathbf{X}_1 = \mathbf{x}) + 1/2 \Pr(T_2 = T_1 | D_2 = 1, D_1 = 0, \mathbf{X}_2 = \mathbf{X}_1 = \mathbf{x}). \quad (1)$$

We adjust for the ties by adding half of the probability that a case and a control has the same test results.

Like in Dodd and Pepe (2003), we assume that AUC_x takes the following generalized linear form:

$$\nu(\mathbf{x}) \equiv AUC_x = g^{-1}(\mathbf{x}^T \boldsymbol{\theta}^*), \quad (2)$$

where $g(\cdot)$ is a monotone link function (known), and $\boldsymbol{\theta}^*$ is the vector of parameters of interest. A convenient choice of the link function may be a logit or probit function. This model makes assumptions on the distribution of the difference of test results between a case and a control with the same covariates: when $D_2 = 1$, $D_1 = 0$ and $\mathbf{X}_2 = \mathbf{X}_1 = \mathbf{x}$, we have $h(T_2 - T_1) = \mathbf{x}^T \boldsymbol{\theta}^* + \epsilon$, where h is some unknown monotone transformation, and ϵ follows the distribution g^{-1} with mean 0.

The binormal ROC curve is a special case of (2), which can be seen from the following example. Suppose the test results for the cases and the controls both follow normal distributions after an unknown monotone transformation h , that is,

$$[h(T) | D=d, \mathbf{X}=\mathbf{x}] \sim \mathcal{N}(\mu(d, \mathbf{x}), \sigma^2(d, \mathbf{x})) \quad (3)$$

where $\mu(d, \mathbf{x})$ and $\sigma^2(d, \mathbf{x})$ are the conditional mean and variance of $h(T)$ as a function of d and $\mathbf{x}$. It is easily verified that the covariate-specific ROC curve is

$$ROC_x(t) = \Phi \left[\frac{\sigma(0, \mathbf{x})}{\sigma(1, \mathbf{x})} \Phi^{-1}(t) + \frac{\mu(1, \mathbf{x}) - \mu(0, \mathbf{x})}{\sigma(1, \mathbf{x})} \right],$$

which takes the “binormal” form with Φ being the standard normal distribution function. The AUC_x could be written explicitly as

$$v(x) \equiv AUC_x = \Phi\left(\frac{a}{\sqrt{b}}\right). \quad (4)$$

where $a = \mu(1, \mathbf{x}) - \mu(0, \mathbf{x})$ and $b = \sigma^2(1, \mathbf{x}) + \sigma^2(0, \mathbf{x})$. Hence if $\mu(d, \mathbf{x})$ is linear in $\mathbf{x}$ and $\sigma^2(d, \mathbf{x})$ is constant in $\mathbf{x}$, then $v(\mathbf{x})$ takes the generalized linear form (2) with link function $g^{-1} = \Phi$. The AUC structure (2) is both parsimonious and flexible, as analogous to generalized regression models for the mean. One may also consider link functions other than Φ .

The definition of AUC_x in (1) restricts the comparison of the test results for subjects in the same covariates' level. However, if some of the covariates are continuous, there may not exist any pairs of a case and a control with the identical covariates value. Hence the estimation of $v(\mathbf{x})$ only based on (2) is not feasible. This prompts us to extend the AUC structure (2) a little bit and allows the comparison of cases and controls with different covariates. Let $\xi(\mathbf{x}, \mathbf{y})$ be the probability of correctly ordering a case with covariates $\mathbf{x}$ and a control with covariates $\mathbf{y}$, i.e.,

$$\begin{aligned} \xi(\mathbf{x}, \mathbf{y}) &= \Pr(T_2 > T_1 | D_2 = 1, D_1 = 0, X_2 = \mathbf{x}, X_1 = \mathbf{y}) \\ &+ 1/2 \Pr(T_2 = T_1 | D_2 = 1, D_1 = 0, X_2 = \mathbf{x}, X_1 = \mathbf{y}). \end{aligned}$$

Under the transformed location-scale model, one can verify the similar result as in (4):

$$\xi(\mathbf{x}, \mathbf{y}) = \Phi\left(\frac{a^*}{\sqrt{b^*}}\right)$$

with $a^* = \mu(1, \mathbf{x}) - \mu(0, \mathbf{y})$ and $b^* = \sigma^2(1, \mathbf{x}) + \sigma^2(0, \mathbf{y})$. Hence a flexible and parsimonious structure of $\xi(\mathbf{x}, \mathbf{y})$ can also be assumed:

$$\xi(\mathbf{x}, \mathbf{y}) = g^{-1}(\mathbf{W}^T \boldsymbol{\theta}), \quad (5)$$

where a natural choice of $\mathbf{W}$ is $\mathbf{W} = (1, \mathbf{x}^T, \mathbf{y}^T)^T$. Obviously, (5) is a special case of (2), because $\xi(\mathbf{x}, \mathbf{x}) = v(\mathbf{x})$.

2.2 Weighted Estimating Equations

Denote $I_{ij} \equiv I(T_i > T_j) + 1/2 I(T_i = T_j)$, where $I(\cdot)$ is the indicator function. When the disease status is observed for every subject, we can construct the full data estimating equations in the form of U statistics. Note that

$$E(I_{ij} | D_i = 1, D_j = 0, \mathbf{X}_i = \mathbf{X}_j) = v(\mathbf{X}_i).$$

Under (2), when the covariates are all discrete, we could consider those pairwise comparisons I_{ij} with $D_i = 1$, $D_j = 0$ and $\mathbf{X}_i = \mathbf{X}_j$. Then the generalized estimating equations for $\boldsymbol{\theta}^*$ are given by

$$0 = \sum_{i \neq j} \left(\frac{\partial v(\mathbf{X}_i)}{\partial \theta^T} \right)^T (\Omega_{ij}^*)^{-1} [I_{ij} - v(\mathbf{X}_i)] D_i (1 - D_j) I(\mathbf{X}_i = \mathbf{X}_j), \quad (6)$$

where $\Omega_{ij}^* = \text{var}(I_{ij} | D_i = 1, D_j = 0, \mathbf{X}_i = \mathbf{X}_j) = v(\mathbf{X}_i)(1 - v(\mathbf{X}_i))$.

When the covariates are continuous, however, the above estimating equations (6) become degenerate, because $I(\mathbf{X}_i = \mathbf{X}_j)$ is always equal to 0. A solution is to construct the estimating equations from the extended AUC form (5). We write $\xi_{ij} \equiv \xi(\mathbf{X}_i, \mathbf{X}_j)$ for short. By noting that

$$E(I_{ij} | D_i = 1, D_j = 0, \mathbf{X}_i, \mathbf{X}_j) = \xi(\mathbf{X}_i, \mathbf{X}_j),$$

we obtain the estimating equations for θ by considering all the pairwise comparison I_{ij} with $D_i = 1$ and $D_j = 0$ as follows:

$$0 = \sum_{i \neq j} U_{ij}^{(FD)} = \sum_{i \neq j} \left(\frac{\partial \xi_{ij}}{\partial \theta^T} \right)^T \Omega_{ij}^{-1} (I_{ij} - \xi_{ij}) D_i (1 - D_j), \quad (7)$$

where $\Omega_{ij} = \text{var}(I_{ij} | D_i = 1, D_j = 0, \mathbf{X}_i, \mathbf{X}_j) = \xi_{ij}(1 - \xi_{ij})$. Note that the full data estimating equations (7) are the classic score equations for binary regression with weight $D_i(1 - D_j)$. The point estimate of θ could be obtained by the binary regression, but the variance estimator needs to be modified to address the cross-correlation of $U_{ij}^{(FD)}$.

Let $\rho_i = \Pr(D_i = 1 | T_i, \mathbf{X}_i)$ be the disease probability, and $\pi_i = \Pr(V_i = 1 | T_i, \mathbf{X}_i, D_i)$ be the verification probability. The disease and verification probabilities often need to be estimated from the data. With the estimated ρ_i and π_i , we may either impute the missing disease status or perform a weighted analysis. Now suppose we have the estimated $\hat{\rho}_i$ and $\hat{\pi}_i$, and we will discuss the estimation of the two probabilities in the next subsection. We propose four types of weighted estimating functions that correct for the verification bias. These estimators work for both MAR and NMAR assumptions. The only difference lies in estimation of the disease and verification models, which we will see in Section 2.3.

The first approach is full imputation (FI), which replaces all the disease status D_i with the estimated probability $\hat{\rho}_i$ in the estimating functions (7):

$$U^{(FI)} = \sum_{i \neq j} U_{ij}^{(FI)} = \sum_{i \neq j} \left(\frac{\partial \xi_{ij}}{\partial \theta^T} \right)^T \Omega_{ij}^{-1} (I_{ij} - \xi_{ij}) \hat{\rho}_i (1 - \hat{\rho}_j). \quad (8)$$

The imputation is performed on every subject regardless of the verification status. So the FI estimator is highly sensitive to the correct specification of the disease model.

Another imputation method is mean score imputation (MSI), which only imputes the missing D_i with $\rho_{0i} \equiv \Pr(D_i = 0 | T_i, \mathbf{X}_i, V_i = 0)$ but keeps the observed ones. The estimated

version, $\hat{\rho}_{0i}$, can be expressed as $\hat{\rho}_{0i} = \frac{(1 - \hat{\pi}_{1i})\hat{\rho}_i}{(1 - \hat{\pi}_{1i})\hat{\rho}_i + (1 - \hat{\pi}_{0i})(1 - \hat{\rho}_i)}$, where $\hat{\pi}_{di} \equiv \Pr(V_i = 1 | T_i, \mathbf{X}_i, D_i = d)$ for $d = 0, 1$. If MAR assumption holds, ρ_{0i} is equal to ρ_i . Let $D_{MSI,i} \equiv V_i D_i + (1 - V_i) \rho_{0i}$ and $\hat{D}_{MSI,i} \equiv V_i D_i + (1 - V_i) \hat{\rho}_{0i}$. The MSI estimating functions are as follows:

$$\mathbf{U}^{(MSI)} = \sum_{i \neq j} \mathbf{U}_{ij}^{(MSI)} = \sum_{i \neq j} \left(\frac{\partial \xi_{ij}}{\partial \theta^T} \right)^T \Omega_{ij}^{-1} (I_{ij} - \xi_{ij}) \widehat{D}_{MSI,i} (1 - \widehat{D}_{MSI,j}). \quad (9)$$

The MSI estimator still requires the disease model to be correctly specified, so that $\widehat{D}_{MSI,i}$ mimics D_i closely. With mis-specified disease probability $\widehat{\rho}_i$, we expect the MSI estimator to work better than the FI estimator. This is because for the MSI estimator, the incorrect imputation model only affects the subjects with missing disease status, while for the FI estimator, the imputation is performed on every subject.

The third method is inverse probability weighting (IPW). Only the complete data are included in the estimating functions that are weighted by the inverse of the sampling probabilities:

$$\mathbf{U}^{(IPW)} = \sum_{i \neq j} \mathbf{U}_{ij}^{(IPW)} = \sum_{i \neq j} \left(\frac{\partial \xi_{ij}}{\partial \theta^T} \right)^T \Omega_{ij}^{-1} (I_{ij} - \xi_{ij}) D_i (1 - D_j) \frac{V_i V_j}{\pi_i \pi_j}. \quad (10)$$

The IPW estimator requires $\widehat{\pi}_i$ to be estimated consistently.

The fourth method is doubly robust (DR) estimator, which makes use of both $\widehat{\rho}_i$ and $\widehat{\pi}_i$. Let $D_{DR,i} \equiv V_i D_i / \pi_i + (1 - V_i) \rho_{0i}$ and $\widehat{D}_{DR,i} \equiv V_i D_i / \widehat{\pi}_i + (1 - V_i / \widehat{\pi}_i) \widehat{\rho}_{0i}$. The DR estimating functions are

$$\mathbf{U}^{(DR)} = \sum_{i \neq j} \mathbf{U}_{ij}^{(DR)} = \sum_{i \neq j} \left(\frac{\partial \xi_{ij}}{\partial \theta^T} \right)^T \Omega_{ij}^{-1} (I_{ij} - \xi_{ij}) \widehat{D}_{DR,i} (1 - \widehat{D}_{DR,j}). \quad (11)$$

The DR estimator requires either $\widehat{\rho}_i$ or $\widehat{\pi}_i$ to be consistently estimated, but not necessarily both. It also implicitly requires that $\widehat{\rho}_i$ and $\widehat{\pi}_i$ can be estimated separately. As we will see in Section 2.3, the disease and verification probabilities are estimated separately under MAR assumption. Under NMAR, if one is confident to specify a nonignorable parameter (the log odds ratio of verification for diseased versus healthy subjects), Rotnitzky *et al.* (2009) demonstrated that $\widehat{\rho}_i$ and $\widehat{\pi}_i$ can still be estimated separately, which leads to the doubly robust property for the AUC estimation. However, if one chooses to estimate the nonignorable parameter, Liu and Zhou (2010) showed that the estimation follows by maximizing the joint likelihood of D_i and V_i . Hence we must specify the correct selection model and disease model, and the doubly robust property is sacrificed.

2.3 Estimating Verification and Disease Processes

From the weighted estimating equations (8)–(11), we note that correctly estimating the disease and verification probabilities ρ_i and π_i is crucial to the estimation of θ and hence AUC_x . We discuss the estimation methods for MAR and NMAR verification processes respectively.

Under the MAR verification process, the disease and verification probabilities can be estimated separately, because

$$\begin{aligned} \Pr(D_i=1|T_i, \mathbf{X}_i) &= \Pr(D_i=1|T_i, \mathbf{X}_i, V_i=1), \\ \Pr(V_i=1|T_i, \mathbf{X}_i, D_i) &= \Pr(V_i=1|T_i, \mathbf{X}_i). \end{aligned}$$

Therefore, the disease probability could be estimated by fitting a binary regression of D_i versus T_i and X_i for the verified cases ($V_i = 1$); the verification probability could be estimated by fitting another binary regression of V_i versus T_i and X_i for all the subjects.

With a NMAR verification process, we could implement the likelihood-based estimators as in Liu and Zhou (2010). The observed data likelihood involves both π_j and ρ_j , which can only be estimated together by solving the score equations. The identifiability of the selection model comes from the parametric assumption of ρ_j on the whole population. The validity of the disease and verification models could not be tested nonparametrically. So we recommend to build the parametric models from the scientific point of view, and to determine the plausible models prior to data analysis.

2.4 Asymptotic Normality

Denote η to be the vector of nuisance parameters in the disease and/or verification models. Our parameters of interest is θ , the covariates effect on AUC_x . All the FI, MSI, IPW and DR estimating functions are U-statistics based on pairwise independently and identically distributed (i.i.d) random variables. We sometimes suppress the superscript and write these estimating functions to be $\Sigma_{i \neq j} U_{ij}(\theta, \eta)$. Note that the estimating functions (8), (9), (10) and (11) are actually $\Sigma_{i \neq j} U_{ij}(\theta, \eta)$. Let θ_0 and η_0 be the true value of the unknown parameters. We begin with showing that $U_{ij}(\theta_0, \eta_0)$ has zero expectation in Lemma 1; then the influence function and limit theorem of the U statistic are given in Lemma 2; finally we complete the proof of the asymptotic normality for $\hat{\theta}$

LEMMA 1: Under the MAR assumption,

- A. if the verification model is correctly specified, then the IPW estimating functions $U_{ij}(\theta_0, \eta_0)$ have zero expectation;
- B. if the disease model is correctly specified, then the FI and MSI estimating functions $U_{ij}(\theta_0, \eta_0)$ both have zero expectation;
- C. if either the verification model or the disease model is correctly specified, then the DR estimating functions $U_{ij}(\theta_0, \eta_0)$ have zero expectation.

Under the NMAR situation,

- D. if both the verification and the disease model are correctly specified, then the IPW, FI, MSI and DR estimating functions $U_{ij}(\theta_0, \eta_0)$ have zero expectation.

The proof of the above lemma is given in the Web Appendix A. Let

$$S_{ij}(\theta, \eta) \equiv (U_{ij}(\theta, \eta) + U_{ji}(\theta, \eta)) / 2.$$

So we have $\Sigma_{i \neq j} U_{ij}(\theta, \eta) = \Sigma_{i \neq j} S_{ij}(\theta, \eta)$ is symmetric in addition. Denote the U-statistic

$$S_n(\theta, \eta) = \frac{1}{n(n-1)} \Sigma_{i \neq j} S_{ij}(\theta, \eta), \text{ the central limit theorem of } S_n \text{ is stated as below:}$$

LEMMA 2: If $\text{var}(S_{ij})$ is finite, then

$$\sqrt{n}(S_n - e) = \frac{1}{\sqrt{n}} \sum_i E[S_{ij}|O_i] + o_p(1) \xrightarrow{d} N(0, 4\sigma_1^2),$$

where $e = ES_{12}(\theta, \eta)$, $\sigma_1^2 \equiv \text{var}(E[S_{12}|\mathbf{O}_1])$ and $\mathbf{O}_i = (V_i, V_i D_i, T_i, X_i)$ is the observed data for the i th subject.

This lemma is a special case of Theorem 6.1.2 of Lehmann (1999). We skip the proof here. With the influence function of S_n in Lemma 2, we are able to analyze the variability of $\hat{\theta}$ using Taylor expansion. As shown in Section 2.3, the nuisance parameters could be estimated from either score equations, or generalized estimating equations, denoted by $\mathbf{B}(\boldsymbol{\eta}) \equiv \Sigma_i \mathbf{B}_i(\boldsymbol{\eta})$. Under the regularity conditions stated in the Web Appendix A, the asymptotic normality results are shown in Theorem 1.

THEOREM 1: Suppose the regularity conditions C1–C6 hold, then

$$\sqrt{n}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0, \Omega)$$

as $n \rightarrow \infty$.

The explicit formula for Ω , as well as the proof of the above theorem are also given in the Web Appendix A. From the proof, we note that the variability of $\hat{\theta}$ comes from two sources: one is the variability due to the U-statistic; the other is the variability due to estimating $\boldsymbol{\eta}$.

3. Estimation for Covariate-Adjusted AUC (AAUC)

3.1 Definition

The AUC_x identifies the risk factors that affect the diagnostic accuracy of the test. However, the policy makers may be more interested in summarizing the overall accuracy across the whole patient population. Such a summary measure was recently proposed by Janes and Pepe (2009). They came up with the covariate-adjusted ROC (AROC) curve, which is a weighted average of all possible covariate-specific ROC curves. In this section, we discuss the estimation method of the area under AROC curve, when verification bias problem is present.

As shown in Janes and Pepe (2009), the AROC curve can be written as a weighted average of the covariate-specific ROC curves:

$$AROC(t) = \int_{-\infty}^{+\infty} ROC_x(t) dF_{X|D=1}(x). \quad (12)$$

The AROC curve at t is interpreted as the average sensitivity when the covariate-specific decision thresholds are chosen to ensure the specificity to be $1 - t$ in each subpopulation. We can also consider the covariate-adjusted AUC, or AAUC:

$$\begin{aligned} AAUC &= \int_0^1 AROC(t) dt \\ &= \int_0^1 \int_{-\infty}^{+\infty} ROC_x(t) dF_{X|D=1}(x) dt \\ &= \int_{-\infty}^{+\infty} AUC_x dF_{X|D=1}(x), \end{aligned}$$

which suggests that AAUC is also a weighted average of AUC_x . The AAUC measures the probability that a *randomly selected case* ($D = 1$) has a greater test result than a *matched control* ($D = 0$).

3.2 Estimation Procedures

With the AUC_x estimator in the previous section, we only need to estimate the covariate distribution for the cases, $F_{X|D=1}(x)$. With the disease status observed for every patient, an empirical estimator is given by:

$$\widehat{F}_{X|D=1}(x) = \frac{\sum_i I(X_i \leq x) D_i}{\sum_i D_i},$$

which is a step function with jump of size $1/\sum_i D_i$ at every data point X_i with $D_i = 1$. Since D_i is missing for some of the patients, we could estimate $F_{X|D=1}(x)$ using FI, MSI, IPW or DR approaches:

$$\widehat{F}_{X|D=1}^{(est)}(x) = \frac{\sum_i I(X_i \leq x) \widehat{D}_i}{\sum_i \widehat{D}_i},$$

where $\widehat{D}_i$ is some version of “estimated” disease status. With similar idea for the estimation of AUC_x , the four different estimators for $\widehat{D}_i$ are available:

$$\widehat{D}_i^{(FI)} = \widehat{\rho}_i,$$

$$\widehat{D}_i^{(MSI)} = V_i D_i + (1 + V_i) \widehat{\rho}_i,$$

$$\widehat{D}_i^{(IPW)} = \frac{V_i}{\pi_i} D_i,$$

$$\widehat{D}_i^{(DR)} = \frac{V_i}{\pi_i} D_i + \left(1 - \frac{V_i}{\pi_i}\right) \widehat{\rho}_i.$$

Therefore, the $AAUC$ is estimated by

$$\begin{aligned} \widehat{AAUC} &= \int_{-\infty}^{+\infty} \widehat{AUC}_x d\widehat{F}_{X|D=1}^{(est)}(x) \\ &= \frac{1}{\sum_i \widehat{D}_i} \sum_i \widehat{AUC}_{x_i} \widehat{D}_i. \end{aligned} \quad (13)$$

As stated in Theorem 1 and its proof, the AUC_x estimator is asymptotically linear. As $AAUC$ is merely a weighted average of AUC_x , the influence function of the $AAUC$ estimator can also be derived. We write $AAUC$ and $\widehat{AAUC}$ to be ν and $\widehat{\nu}$ for short, and it follows from (13) that

$$0 = \sum_i R_i(\widehat{\theta}, \widehat{\eta}, \widehat{\nu}) \equiv \sum_i (\widehat{AUC}_{x_i} - \widehat{\nu}) \widehat{D}_i.$$

Note that $\widehat{AUC}_{x_i}$ is a function of $\widehat{\theta}$ and $\widehat{\eta}$, and $\widehat{D}_i$ is a function of $\widehat{\eta}$.

THEOREM 2. If the regularity conditions C1–C6 hold, $\widehat{AAUC}$ is asymptotically linear:

$$\sqrt{n}(\widehat{AAUC} - AAUC) = \frac{1}{\sqrt{n}} \sum_i \Psi_i + o_p(1),$$

where the form of Ψ_i is given in the Web Appendix A. Consequently, the asymptotic variance for $\widehat{AAUC}$ is given by $\text{var}(\Psi_1)$.

Theorem 2 is proved by Taylor expansion on $R_i(\widehat{\theta}, \widehat{\eta}, \widehat{\nu})$ around the true values θ_0 , η_0 and ν . The detailed proof is given in the Web Appendix A.

4. Simulation Studies

We conducted two sets of simulation studies to examine the finite sample performance for the AUC_x and $AAUC$ estimators in this section. The first simulation study assumed the MAR verification process, and the second examined NMAR situation. More extensive simulation studies were reported in the Web Appendix B, which examined the different sample size, AUC , disease prevalence and verification proportion.

4.1 Simulation One: MAR Verification Process

We generated two covariates X_1 and X_2 from $Bernoulli(0.5)$ and $U(-1, 1)$ distributions, respectively. The true disease status D was generated from $D|X_1, X_2 \sim Bernoulli(p)$, where $\text{logit}(p) = -1.4 + 0.5X_1 + 0.8X_2$. The test result T was generated from a location-scale model, $T|D, X_1, X_2 \sim N(\mu, \sigma^2)$, where $\mu = 1 + 0.4D + 0.2X_1 + 0.7X_2 + X_1D + 0.5X_2D$ and $\sigma = 0.8D + 1.2(1 - D)$. The verification indicator V was generated from $V|T, X_1, X_2 \sim Bernoulli(\pi)$, where $\text{logit}(\pi) = -1.2 + T + 0.6X_1 + 1.2X_2$. The disease prevalence was about 25% and the verification proportion was about 57%, which was similar to our real example in the next section. The true AUC_x was given by $\Phi(\theta_0^* + \theta_1^*X_1 + \theta_2^*X_2)$, where $(\theta_0^*, \theta_1^*, \theta_2^*) = (0.277, 0.693, 0.453)$.

Since the disease probability, $p = \Pr(D=1|T, X_1, X_2)$, is jointly determined by $D|X_1, X_2$ and $T|D, X_1, X_2$, we show in the Web Appendix A that $\text{logit}(p)$ is indeed a quadratic form of (T, X_1, X_2) under our data generating procedure. The correct disease model we used here was a logistic regression with main effects and pairwise interactions of T, X_1 , and X_2 , as well as the quadratic terms of T and X_2 . We also examined the mis-specified disease and verification models: the mis-specified disease model ignored the quadratic term of X_2 and the interaction between X_1 and X_2 , while the mis-specified verification model ignored the effect of X_2 . We set the sample size to be 1000 and repeated the simulation for 1000 times. Table 1 displays the results for estimating θ^* , AUC_x and $AAUC$, where a total of 12 estimators were calculated:

- (1) Ideal: full data analysis, i.e., the true disease status was observed for every subject.
- (2) CC: complete-case analysis, i.e., all the non-verified subjects were removed.
- (3) IPW₁: the IPW estimator with correct verification model.
- (4) IPW₂: the IPW estimator with incorrect verification model.

- (5) FI_1 : the FI estimator with correct disease model.
- (6) FI_2 : the FI estimator with incorrect disease model.
- (7) MSI_1 : the MSI estimator with correct disease model.
- (8) MSI_2 : the MSI estimator with incorrect disease model.
- (9) DR_1 : the DR estimator with correct disease and verification models.
- (10) DR_2 : the DR estimator with incorrect disease model and correct verification model.
- (11) DR_3 : the DR estimator with correct disease model and incorrect verification model.
- (12) DR_4 : the DR estimator with incorrect disease and verification models.

The estimated coefficients θ^* were shown on the top of Table 1. The ideal estimator used the full data with all D_j available, and hence is unfeasible in practice. It just served as a reference to indicate the amount of information gained by observing the values of the missing data. The CC estimator was biased as we expected, because the complete data were no longer a random sample from the population. The IPW_1 , FI_1 , MSI_1 , DR_1 estimators all performed well under correct model specification. When either the disease or verification model was wrong, the DR_2 and DR_3 estimators still worked as good as the DR_1 estimator in terms of unbiasedness, efficiency and coverage rate. It is noted by a referee that the DR_3 estimator yielded less bias than DR_1 . We examined the fitted verification probabilities for the correct and incorrect model in one simulated data set. The estimated “weight” function $\widehat{D}_{DR,i} (1 - \widehat{D}_{DR,j})$ in (11) could be large if both $\widehat{\pi}_i$ and $\widehat{\pi}_j$ are small. We found that the verification probabilities in DR_1 is more scattered within (0, 1) than those in DR_3 . Therefore, DR_1 generally has more large weights than DR_3 , and is less stable in finite samples. This would not be a serious issue as the sample size increases, as both estimators are consistent. In fact, in an unreported simulation study with $n = 2000$, we observed less bias for the DR_1 estimator. It would be interesting future research to investigate the impact of “extreme” weights in the DR and IPW estimators.

The FI_2 , MSI_2 , IPW_2 and DR_4 estimators for θ^* were all biased, but the latter three still have reasonable CI coverage (between 87% and 94%), compared to that of the CC estimator (between 39% and 91%). This suggests that when the model mis-specification is moderate, the proposed MSI_2 , IPW_2 and DR_4 estimators may still be used in practice to correct for some verification bias. We also noted that FI_2 estimator was the most biased, because the disease status was imputed for every subject using the wrong disease model. As a comparison, DR_4 and MSI_2 estimators provided better protection against incorrect disease model.

The performance of AUC_x and AAUC estimators was summarized in the bottom of Table 1. The proposed IPW_1 , FI_1 , MSI_1 , DR_1 , DR_2 , DR_3 estimators still yielded good results. It appeared that $AUC_{(1;0.5)}$ was consistent even for MSI_2 and DR_4 estimators, but this is just because the negatively biased $\widehat{\theta}_1^*$ and $\widehat{\theta}_2^*$ cancelled out with the positively biased $\widehat{\theta}_0^*$. It was somewhat surprising that the AAUC estimators were not very sensitive to model mis-specification except for the IPW estimator. This is probably because the bias in estimating AUC_x could be in either direction depending on the choice of x , and the bias may cancel out in computing AAUC as the weighted average of AUC_x . As for the relative efficiency of the proposed estimators, we found that imputation (FI and MSI) led to the best precision, and IPW estimator had the least precision.

To examine more serious model mis-specifications, we conducted a separate simulation study. We use different (wrong) working disease and verification models: the disease model only included the main effects of T , X_1 and X_2 ; the verification model only included the main effects of X_1 and X_2 , but not T . The results for FI_2 , MSI_2 , and IPW_2 estimators are shown in Table 5 of the Web Appendix. The DR_4 estimator did not converge for 20–30% of data generations, so the results were not reported. An intuitive explanation of non-convergence is that some of the “weights” in equation (11), $\widehat{D}_{DR,i}(1 - \widehat{D}_{DR,j})$, are negative. The negative weights in a regression force the fitted regression line to be as far from the corresponding data points as possible. There is no guarantee that such estimating equations have any solution, as the quasi likelihood function may be no longer concave. On the contrary, the weights of estimating equations in (8)–(10) are all positive, and all the simulations converged for FI_2 , MSI_2 , and IPW_2 estimators. Now the bias for both θ^* and AUC_x was more substantial, and the coverage rate was much lower than 95%. However, the MSI and FI estimators still worked well for estimating $AAUC$.

In practice, if one is confident about the disease model, then MSI and FI estimators are recommended for better power. Otherwise, the DR estimator would be a safe choice that protects some model mis-specification.

4.2 Simulation Two: NMAR Situation

For the NMAR verification process, we generated X_1 , X_2 , D , and T similarly as in the previous subsection. The verification indicator V was generated from $V|T, X_1, X_2, D \sim \text{Bernoulli}(\pi)$, where $\text{logit}(\pi) = -1.5 + T + 0.6X_1 + 1.2X_2 + 2D$. The verification proportion was about 58% in this case. We repeated the simulation for 1000 times. With the sample size being 2000, the NMAR model did not converge for 4.9% generations, so we only reported the results for the converged data sets. In Table 2, we found that the complete-case analysis was seriously biased as we expected. The bias for the proposed estimators was small for the reasonably large sample size. The CI coverage rate was close to 95% nominal level. But as we show in the appendix, when $n = 1000$, the NMAR model could be a bit more biased for θ^* , and more generations (14.6%) did not converge. With smaller sample size, the data do not contain enough information to effectively estimate the nonignorable parameter. The non-convergence occurs if a boundary solution is obtained; even for converged generations, the standard error for the nonignorable parameter might be large, leading to unstable AUC estimators. This nonconvergence issue was also noted by several other authors (Zhou and Castelluccio, 2003; Kosinski and Barnhart, 2003; Liu and Zhou, 2010).

The cost of estimating the NMAR model is the loss of robustness. Therefore, if NMAR verification process is conjectured in practice, one needs to be cautious to specify the correct disease and verification models.

5. Example

We applied the proposed AUC_x and $AAUC$ estimators to an Alzheimer's disease research. We used a data set within the Uniform Data Set (UDS) of National Alzheimer's Coordinating Center (NACC), which came from 32 Alzheimer's Disease Centers throughout North America since 2006. The patients were referred or self-referred for evaluation of possible dementia, or recruited specifically to participate in clinical research. Most patients underwent clinical evaluation and neuropsychological tests for cognitive impairment at enrollment. During the follow-up period, the patients received periodical re-evaluation and cognitive tests. Among these cognitive tests, the mini-mental state examination (MMSE) is a brief 30-point questionnaire test that is widely used to screen for cognitive impairment. In

the progression of dementia, the amnesic mild cognitive impairment (aMCI) is an important transitional stage. Patients with aMCI could still revert to normal, but dementia is generally believed to be irreversible. We are interested to investigate the one-year progression from aMCI to dementia, and find out how well the baseline MMSE score classifies the patients who progressed to dementia and those who do not in one year. The classification ability of MMSE may also depend on patient characteristics, so we used AUC_x to describe the covariate effects, and AAUC to evaluate the overall classification accuracy.

We included the patients who aged over 65 and was diagnosed to be aMCI at their first visit. If a patient made a visit about one year (within the 6–18 months window) after the baseline, his/her cognitive status is observed with $D = 1$ indicating progression to dementia and 0 otherwise. The disease status is missing if the patient only made the baseline visit, or the follow-up visits were all outside the 6–18 months window. This led to about 56.1% disease verification. Within the verified sample, the progression probability was about 24.9%. The covariates we adjusted for in the ROC analysis include age, gender, race, marital status, living situation, stroke, and history of cardiovascular diseases. We also collected other disease history variables, and clinical dementia rating (CDR) sum of boxes as the predictor for the missingness mechanism and the disease model. For simplicity, subjects with missing covariates were excluded, and our final sample for analysis consisted of 2,702 subjects. The list of variables and their summary statistics are shown in the Web Appendix C.

We started with the MAR assumption. Logistic regressions were used for estimating the disease and verification probabilities. The verification model included the MMSE score (T) and its quadratic term, the covariates we adjusted for in the ROC analysis, the disease history information, CDR sum of boxes, interaction between T and stroke, and interaction between T and history of cardiovascular diseases. The disease model included all the main effects in the verification model, as well as the interactions between T and the covariates in the ROC analysis. Under the MAR assumption, the estimated covariate effects are listed in the top of Table 7.

Although the disease and verification models included many risk factors, unobserved confounders may still exist, because (i) the progression of dementia is complicated and not fully understood, and (ii) in this observational study, the missingness could be due to various reasons. So we would recommend the DR estimator in this example, which protects the model mis-specification. The DR estimator showed that MMSE has significantly worse classification accuracy for patients with stroke (probit $AUC_x = -0.433$, 95% CI: $-0.786, 0.080$), and significantly better accuracy for patients with more than 17 years of education (probit $AUC_x = 0.259$, 95% CI: $0.024, 0.494$). Age and cardiovascular history are both marginally significant in affecting the accuracy. The magnitude of the coefficients are interpreted in the probit scale. For example, the IPW estimator showed that, with 10 years increase in age, the probit AUC_x decreased by 0.132, with 95% CI being $(0.005, 0.259)$. We also noted that the imputation-based estimators, especially FI estimator, is a bit deviated from others, suggesting that the progression from aMCI to dementia might be very complicated, and not well estimated by the working disease model. The AAUC was estimated to be about 0.64. This implies that, although MMSE score has some classification ability to predict the progression to dementia, it is less than a satisfactory marker. Further research is motivated to find a combination of several tests as a better marker. The complete-case analysis was not too different from the DR estimator, because the verification process was not so strongly affected by the covariates. In Figure 1 of the Web Appendix, we plotted out the AUC_x as a function of age, according to the DR estimator. Other covariates are fixed as a white male, single, living alone, with a history of cardiovascular diseases, no stroke, and over 17 years of education. We can observe the decreasing trend, but it is inconclusive at the 5% significance level.

The sicker patients might be more likely to miss a follow-up visit, and hence have missing D . Even though the verification model included MMSE and CDR sum of boxes as the measurement of disease severity, there may still be unobserved attributes associated with the disease verification. Therefore, we also investigated the NMAR situation in our data set as a sensitivity analysis, and the results are presented in the bottom half of Table 7. The IPW estimator was close to the estimators under MAR assumption. The MSI and FI estimators differed from the MAR estimators, perhaps due to the unsatisfactory disease model.

6. Discussion

This paper tackled the verification bias problem in ROC analysis with covariate adjustment. We specifically focused on AUC_x and AAUC, and proposed several verification bias-corrected estimators. The proposed estimation procedure used the idea that the pairwise comparison of the test result has the conditional expectation that is exactly the AUC_x . To obtain the overall measure, AAUC, we took the weighted average of AUC_x over the covariates distribution of the diseased subjects. The estimating equations for AUC_x and AAUC were constructed, and reweighted in several different fashions. We proved the central limit theorem for AUC_x and AAUC estimators, and derived the analytic form of the asymptotic variances. We noted that this reweighting approach works for both MAR and NMAR verification processes, while the only difference was in estimating the nuisance parameters. Simulation studies were conducted to confirm the good finite-sample performance of the proposed estimators and their analytical variance formula.

The proposed AUC_x estimators make an extension to Dodd and Pepe (2003) to allow for the missing gold standard. Compared to Liu and Zhou (2011), our focus is in the covariate effects on the diagnostic accuracy itself in this paper. Therefore, we choose a summary measure, AUC_x , and directly make assumptions on its form as a function of covariate $\mathbf{x}$, while Liu and Zhou (2011) modeled the test results first and then derived the ROC_x expression. We note that either of the two approaches is not the special case of the other, but have different focuses. When one is more interested to estimate the ROC curve for particular groups of people, the method of Liu and Zhou (2011) would apply; if one wishes to determine directly how the accuracy is affected by covariates, the proposed estimators could answer the question. In a sense, the parameter interpretation in AUC_x is easier and more direct.

The NMAR model requires both the disease and verification models to be correct. While the model assumptions could not be tested nonparametrically using the observed data, we recommend building up plausible models from the researcher's scientific knowledge. In practice, the NMAR model could serve as a sensitivity analysis to supplement the MAR model. The traditional setting of the sensitivity analysis (Rotnitzky, *et al.* 2006) involves fixing the nonignorable parameter at different values, but it remains unclear which value is the most likely. Here we make additional model assumptions from the prior knowledge, and make use of the “model information” to infer about the nonignorable parameter.

A possible weakness of the proposed estimators is the computation complexity. As the pairs of the test results $\{I_{ij}\}$ contribute to the estimating equations, the computation load could be extremely heavy for large data sets. For the sample size $n = 10,000$, the FI, MSI and DR estimators all need to fit a weighted binary regression of sample size $n^2 = 10^8$. The IPW estimator is faster, because only the I_{ij} pairs with $D_i = 1$ and $D_j = 0$ have non-zero weights. If the disease prevalence is 50%, the weighted regression has sample size $(n \times 0.5)^2 = 2.5 \times 10^7$. Generally speaking, the computation load is approximately $O(n^2)$ for the proposed estimators. Another possible limitation of our proposed method is the parametric form of the link function. Although the probit or logit link covers a wide variety of possible AUC

structure, there may be examples that the true link function is non-standard, especially when the test distribution is highly skewed. This motivates for future research on single index models for the AUC.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors would like to thank National Alzheimer's Coordinating Center (NACC) for providing the data for analysis. The authors are grateful to the associate editor and two anonymous referees for their insightful comments, which greatly improved the quality of this paper. The work was supported in part by NIH/NIA grant U01AG016976. This paper does not necessarily represent the findings and conclusions of VA HSR&D. Dr. Xiao-Hua Zhou is presently a Core Investigator and Biostatistics Unit Director at HSR&D Center of Excellence, Department of Veterans Affairs Puget Sound Health Care System, Seattle, WA.

References

- Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society - Series C (Applied Statistics)*. 2005; 54:173–190.
- Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to verification bias. *Biometrics*. 1983; 39:207–215. [PubMed: 6871349]
- Bennett DA, Schneider JA, Aggarwel NT, Arvanitakis Z, Shah RC, Kelly JF, Fox JH, Cochran EJ, Arends D, Treinkman A, Wilson RS. Decision rules guiding the clinical diagnosis of Alzheimer's disease in two community-based cohort studies compared to standard practice in a clinic-based cohort study. *Neuroepidemiology*. 2006; 27:169–176. [PubMed: 17035694]
- Dodd LE, Pepe MS. Semiparametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association*. 2003; 98:409–417.
- Fluss R, Reiser B, Faraggi D, Rotnitzky A. Estimation of the ROC curve under verification bias. *Biometrical Journal*. 2009; 51:475–490. [PubMed: 19588455]
- He H, Lyness ML, McDermott MP. Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine*. 2009; 28:361–376. [PubMed: 18680124]
- Janes H, Pepe MS. Adjusting for covariate effects on classification accuracy using the covariate-adjusted receiver operating characteristic curve. *Biometrika*. 2009; 96:371–382. [PubMed: 22822245]
- Koepsell TD, Chi YY, Zhou XH, Lee WW, Ramos EM, Kukull WA. An alternative method for estimating efficacy of the AN1792 vaccine for Alzheimer disease. *Neurology*. 2007; 69:1868–1872. [PubMed: 17984455]
- Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*. 2003; 59:163–171. [PubMed: 12762453]
- Lehmann, EL. *Elements of Large Sample Theory*. Springer; New York: 1999.
- Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2nd edition. John Wiley; New York: 2002.
- Liu D, Zhou XH. A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach. *Biometrics*. 2010; 66:1119–1128. [PubMed: 20222937]
- Liu D, Zhou XH. Semiparametric estimation of the covariate-specific ROC curve in presence of ignorable verification bias. *Biometrics*. 2011; 67:906–916. [PubMed: 21361890]
- Page JH, Rotnitzky A. Estimation of the disease-specific diagnostic marker distribution under verification bias. *Computational Statistics and Data Analysis*. 2009; 53:707–717. [PubMed: 23087495]

- Rotnitzky A, Faraggi D, Schisterman E. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*. 2006; 101:1276–1288.
- Zhou XH, Castelluccio P. Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. *Journal of Statistical Planning and Inference*. 2003; 115:193–213.

Table 1

The bias (in percentage of the true value), average standard error (SE), empirical standard deviation (SD) for θ^* , AUC_x and $AAUC$ estimators under MAR verification process.

$\theta_0^*=0.277$																$\theta_1^*=0.693$																$\theta_2^*=0.347$															
Estimator	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)																							
Ideal	1.1	7.35	7.06	95.8	0.0	10.36	10.09	95.2	-1.0	9.13	9.11	94.6																																			
CC	-86.1	10.84	10.97	39.7	10.3	13.46	13.53	91.4	43.9	12.63	12.48	77.4																																			
FI ₁	0.6	10.88	11.24	94.8	1.1	12.63	12.69	94.1	-0.8	11.97	12.06	95.0																																			
FI ₂	49.7	10.42	10.81	71.7	-25.1	12.11	12.07	68.9	-55.0	11.26	11.47	57.5																																			
MSI ₁	0.6	10.93	11.31	95.0	0.8	12.78	12.87	94.0	-1.2	12.10	12.10	95.3																																			
MSI ₂	26.8	10.29	10.73	87.0	-9.3	12.11	12.21	90.8	-26.7	11.16	11.17	84.8																																			
IPW ₁	4.3	12.69	14.12	92.2	-0.7	16.18	17.63	92.3	-4.4	15.23	17.01	92.2																																			
IPW ₂	19.2	12.89	13.94	90.1	0.4	16.02	16.89	93.9	30.3	14.96	16.03	87.8																																			
DR ₁	6.9	11.92	13.23	90.6	-1.5	14.83	16.03	93.4	-7.8	13.97	15.66	89.6																																			
DR ₂	7.3	12.30	13.68	90.7	-1.7	15.19	16.50	93.5	-8.1	14.54	16.16	90.2																																			
DR ₃	4.2	11.38	12.30	92.5	-0.6	14.12	14.83	93.8	-3.4	12.63	13.80	91.8																																			
DR ₄	16.4	10.93	11.95	88.4	-2.5	13.93	14.62	93.7	-14.2	12.20	13.35	89.9																																			

$AUC_{(0,0)} = 0.6092$																$AUC_{(1,0.5)} = 0.8737$																$AAUC = 0.7582$															
Estimator	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)																							
Ideal	0.1	2.81	2.70	95.8	-0.1	1.76	1.70	96.5	0.0	1.67	1.63	95.3																																			
CC	-15.4	4.29	4.35	39.7	-2.4	2.20	2.14	83.6	-2.1	2.14	2.12	91.9																																			
FI ₁	-0.0	4.15	4.29	94.8	0.1	1.84	1.80	96.4	-0.1	2.56	2.60	94.4																																			
FI ₂	8.4	3.80	3.93	71.7	-3.4	1.99	1.94	63.0	-0.1	2.50	2.55	94.3																																			
MSI ₁	-0.0	4.17	4.31	95.0	0.0	1.88	1.82	95.8	-0.2	2.57	2.60	95.1																																			
MSI ₂	4.5	3.84	4.00	87.0	-1.0	1.86	1.80	93.6	-0.1	2.55	2.60	94.7																																			
IPW ₁	0.6	4.82	5.36	92.2	-0.2	2.26	2.32	95.2	-0.1	2.71	2.84	93.2																																			
IPW ₂	3.1	4.84	5.22	90.1	2.3	1.95	1.94	83.4	3.4	2.50	2.55	73.8																																			
DR ₁	1.0	4.52	5.02	90.6	-0.3	2.14	2.21	94.4	0.0	2.64	2.82	91.7																																			

Estimator	$AUC_{(0,0)} = 0.6092$						$AUC_{(1,0.5)} = 0.8737$						$AAUC = 0.7582$					
	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)		
DR ₂	1.1	4.65	5.17	90.7	-0.3	2.20	2.28	94.8	0.0	2.66	2.86	91.6						
DR ₃	0.6	4.34	4.68	92.5	-0.1	2.04	2.03	95.6	-0.0	2.59	2.67	93.6						
DR ₄	2.7	4.12	4.50	88.4	-0.0	2.00	1.98	95.8	-0.0	2.58	2.68	93.1						

Table 2

The bias (in percentage of the true value), average standard error (SE), empirical standard deviation (SD) for θ^* , AUC_x and $AAUC$ estimators under NMAR verification process.

$\theta_0^* = 0.277$												
$\theta_1^* = 0.693$												
$\theta_2^* = 0.347$												
Estimator	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)
CC	-144.0	6.99	7.24	0.0	17.6	9.04	9.31	72.1	72.1	8.35	8.54	15.3
FI ₁	-5.9	8.94	8.96	92.8	1.8	8.98	9.13	94.4	4.2	8.69	8.88	93.9
MSI ₁	-6.2	8.98	8.96	93.1	1.6	9.10	9.18	94.7	3.6	8.79	8.97	93.6
IPW ₁	-2.7	10.11	10.21	93.6	-0.2	11.61	12.01	93.9	-0.0	11.13	11.39	93.9

$AUC_{(0,0)} = 0.6092$												
$AUC_{(1,0.5)} = 0.8737$												
$AAUC = 0.7582$												
Estimator	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)	Bias (%)	SE×100	SD×100	Coverage (%)
CC	-25.9	2.76	2.86	0.0	-4.0	1.64	1.71	38.8	-10.0	1.59	1.65	0.5
FI ₁	-1.1	3.44	3.46	92.8	0.0	1.26	1.31	94.6	-1.2	2.85	2.87	91.8
MSI ₁	-1.2	3.46	3.56	93.1	-0.1	1.29	1.34	94.3	-1.2	2.67	2.86	91.4
IPW ₁	-0.6	3.88	3.93	93.6	-0.3	1.60	1.68	94.1	-1.2	2.86	2.94	92.0

Table 3

Estimated covariate effects on AUC_x , AAUC and their associated standard errors for the NACC UDS example.

	<i>DR</i> (MAR)	<i>IPW</i> (MAR)	<i>MSI</i> (MAR)	<i>FI</i> (MAR)
Age (per 10 years)	-0.118 (0.061)	-0.132 (0.065) ^a	-0.106 (0.060)	-0.075 (0.058)
Gender (Male vs. Female)	-0.030 (0.097)	-0.019 (0.100)	-0.058 (0.091)	-0.105 (0.088)
Race (White vs. Non-white)	-0.027 (0.130)	-0.085 (0.132)	-0.054 (0.128)	-0.091 (0.126)
Education 12- years (reference)	-	-	-	-
Education 13-16 years	0.085 (0.110)	0.122 (0.112)	0.066 (0.105)	0.043 (0.100)
Education 17+ years	0.259 (0.120)	0.282 (0.121)	0.232 (0.113)	0.202 (0.108)
Living situation (Alone vs. Others)	0.106 (0.146)	0.035 (0.160)	0.084 (0.139)	0.077 (0.132)
Marital status (Married vs. Others)	0.100 (0.149)	0.023 (0.164)	0.111 (0.142)	0.132 (0.134)
Stroke (Yes vs No)	-0.433 (0.180)	-0.431 (0.191)	-0.403 (0.180)	-0.338 (0.178)
Cardiovascular history (Yes vs No)	0.142 (0.094)	0.188 (0.096)	0.163 (0.087)	0.187 (0.084)
AAUC	0.6371 (0.0163)	0.6398 (0.0166)	0.6372 (0.0164)	0.6379 (0.0164)

	<i>CC</i>	<i>IPW</i> (NMAR)	<i>MSI</i> (NMAR)	<i>FI</i> (NMAR)
Age (per 10 years)	-0.143 (0.064)	-0.137 (0.065)	-0.144 (0.057)	-0.125 (0.052)
Gender (Male vs. Female)	-0.012 (0.098)	-0.017 (0.098)	-0.025 (0.091)	-0.029 (0.085)
Race (White vs. Non-white)	-0.077 (0.132)	-0.074 (0.131)	-0.017 (0.123)	0.015 (0.123)
Education 12- years (reference)	-	-	-	-
Education 13-16 years	0.116 (0.112)	0.119 (0.111)	0.066 (0.103)	0.023 (0.098)
Education 17+ years	0.271 (0.120)	0.275 (0.120)	0.189 (0.111)	0.063 (0.104)
Living situation (Alone vs. Others)	0.003 (0.158)	0.010 (0.158)	0.027 (0.138)	-0.089 (0.120)
Marital status (Married vs. Others)	0.001 (0.161)	-0.001 (0.161)	0.038 (0.141)	-0.103 (0.121)
Stroke (Yes vs No)	-0.425 (0.188)	-0.430 (0.185)	-0.360 (0.181)	0.016 (0.204)
Cardiovascular history (Yes vs No)	0.190 (0.095)	0.200 (0.093)	0.159 (0.094)	0.016 (0.085)
AAUC	0.6388 (0.0166)	0.6374 (0.0161)	0.6291 (0.0159)	0.6229 (0.0151)

^a Coefficients in bold are significant at $p < 0.05$.