Published in final edited form as:
Biometrics. 2011 June ; 67(2): 559–567. doi:10.1111/j.1541-0420.2010.01487.x.

Robust Estimation of Area Under ROC Curve Using Auxiliary Variables In the Presence of Missing Biomarker Values

Qi Long[1,*], **Xiaoxi Zhang**[2], **Brent A. Johnson**[1]

[1]Department of Biostatistics and Bioinformatics, Emory University, Atlanta, GA 30322, U.S.A.

[2]Pfizer Inc., New York, NY 11017, U.S.A.

Summary:

In medical research, the receiver operating characteristic (ROC) curves can be used to evaluate the performance of biomarkers for diagnosing diseases or predicting the risk of developing a disease in the future. The area under the ROC curve (AUC), as a summary measure of ROC curves, is widely utilized, especially when comparing multiple ROC curves. In observational studies, the estimation of the AUC is often complicated by the presence of missing biomarker values, which means that the existing estimators of the AUC are potentially biased. In this article, we develop robust statistical methods for estimating the ROC AUC and the proposed methods use information from auxiliary variables that are potentially predictive of the missingness of the biomarkers or the missing biomarker values. We are particularly interested in auxiliary variables that are predictive of the missing biomarker values. In the case of missing at random (MAR), i.e., missingness of biomarker values only depends on the observed data, our estimators have the attractive feature of being consistent if one correctly specifies, conditional on auxiliary variables and disease status, either the model for the probabilities of being missing or the model for the biomarker values. In the case of missing not at random (MNAR), i.e., missingness may depend on the unobserved biomarker values, we propose a sensitivity analysis to assess the impact of MNAR on the estimation of the ROC AUC. The asymptotic properties of the proposed estimators are studied and their finite sample behaviors are evaluated in simulation studies. The methods are further illustrated using data from a study of maternal depression during pregnancy.

Keywords

Area under the curve; Biomarker; Doubly robust estimators; Missing at random; Missing not at random; Receiver operating characteristic curve; Sensitivity analysis

1. Introduction

The receiver operating characteristic (ROC) curve plots the fraction of true positives (sensitivity) against the fraction of false positives (1–specificity) as the discrimination

[*] qlong@emory.edu.

Supplementary Materials
Web Appendices referenced in Sections 2 and 3 are available under the Paper Information link at the Biometrics website http://www.biometrics.tibs.org.

threshold (e.g., of a biomarker for a disease) is varied, and it is often used to evaluate the performance of biomarkers for diagnosing diseases or predicting the risk of developing diseases in the future. It was originally developed for the analysis of signal detection (Green and Swets, 1966) and was first used in medicine for the assessment of imaging devices (Zweig and Campbell, 1993). In medical studies, summary measures of ROC curves are often used and they are particularly powerful when comparing several ROC curves. The most widely used summary measure is the area under the ROC curve (ROC AUC) (Bamber, 1975). The ROC AUC is bounded between 0.5 and 1, and has the interpretation of the probability of a randomly selected observation from the diseased (non-diseased) population having a higher biomarker value than that from the non-diseased (diseased) population. Therefore, a large AUC value represents good separation in the biomarker values between the diseased and non-diseased populations. In particular, a perfect test would achieve an AUC of 1.0, whereas an uninformative test would have an AUC of 0.5. A wealth of literature has been developed for this type of research (Pepe (2003) and references therein).

In practice, the biomarker value may be missing for some subjects, especially in observational studies. Take for example a self-rated mental illness score collected from pregnant women in a psychiatric study, where the disease of interest is the presence (or absence) of a major depressive episode throughout pregnancy (see Section 4 for more details). Since the biomarker score is self-rated, it is possible that some subjects did not complete the self-evaluation and hence the score is missing. In such studies, additional variables including demographic and baseline variables are often available, which are referred to as auxiliary variables. While these variables are not of primary interest themselves, they are potentially predictive of the missingness of the biomarker value or the value itself, and can be incorporated in a data analysis to improve its robustness and/or efficiency. If an auxiliary variable is predictive of missingness but independent of the missing values, then using it in an analysis will not affect the results. Thus, we are interested in auxiliary variables that are predictive of the missing values, especially if they are also predictive of the missingness.

As with the general setting discussed in Little and Rubin (2002) and references therein, a naive analysis that only uses complete observations may lead to bias and loss of efficiency in the estimation of the ROC AUC. First, when the biomarker is missing completely at random (MCAR), i.e., the missingness does not depend on either observed or unobserved data, the naive analysis is valid but is not efficient. Second, when the biomarker is missing at random (MAR), i.e., the missingness is conditionally independent of the missing data given the observed data, the naive analysis is biased and other methods, e.g., inverse-weighted (IW) methods, can be extended for consistent estimation. IW methods weight each complete case by the inverse of the probability of observing the biomarker value. Despite its conceptual simplicity, IW methods have limitations. Most notably, IW methods are not efficient and are subject to bias if one misspecifies the model for the missingness. Alternatively, one can extend the methods that are doubly robust and more efficient (Robins et al., 1994; Scharfstein et al., 1999) for estimating the ROC AUC. In the case of missing not at random (MNAR), i.e., missingness depends on unobserved biomarker values even after conditioning on the observed data, it is common practice to conduct sensitivity analysis (Zhou, 1994; Rotnitzky and Robins, 1997; Scharfstein et al., 1999; Kosinski and Barnhart, 2003). In all cases, auxiliary variables can be used to potentially reduce bias and improve efficiency when associated with the probability of missing and the value of biomarkers, or simply improve efficiency when only associated with the value of biomarkers.

We confine the scope of this paper to the case where the disease status is always confirmed and a set of auxiliary variables are fully observed but the biomarker values are missing for some subjects, and we are interested in estimating the ROC AUC. Our setting is to be distinguished from the existing research on verification bias (Zhou, 1993, 1998; Rotnitzky et al., 2006; Fluss et al., 2009). In the presence of verification bias, the biomarker values are always observed whereas the true disease status is only verified for a non-random sample of the population of interest, e.g., the selection for testing may depend on the disease status or other variables. In particular, Rotnitzky et al. (2006) extended the doubly robust method developed in Rotnitzky and Robins (1997) to the estimation of the ROC AUC in the presence of verification biases. As a result of different problem setups (i.e. biomarker values missing vs. disease status unconfirmed for a subset of subjects), there are important differences between our work and theirs. In our setting, a working model on biomarker values, which can be continuous or categorical, is utilized, whereas a working model on the presence (or absence) of the disease, a binary variable, was utilized in Rotnitzky et al. (2006); consequently, our methods require modeling of the conditional distribution of biomarker values. Furthermore, we study and compare parametric and nonparametric approaches for estimating this conditional distribution and discuss two types of MAR assumptions, which have different implications on the estimation of AUC.

The remainder of the article is organized as follows. In Section 2, we describe the proposed estimators and their theoretical properties under MAR and propose a sensitivity analysis under MNAR. In Section 3, we evaluate the finite sample performance of the proposed estimators through simulations. In Section 4, we apply the proposed methods to a psychiatric study of maternal depression during pregnancy. We conclude with a discussion in Section 5.

2. Methodology

Suppose that a random sample of n subjects is selected from a population of interest to evaluate the performance of a diagnostic or predictive test using a biomarker. Each subject i, $i = 1, \dots, n$, is classified into one of two groups, diseased ($D_i = 1$) or non-diseased ($D_i = 0$), based on a gold standard. For each subject i, denote the biomarker value by X_i, which is used to diagnose or predict the disease status (D_i). X_i is not observed in a subset of the subjects, and let δ_i denote the missing indicator for X_i, i.e., $\delta_i = 1$ when X_i is observed and $\delta_i = 0$ if X_i is missing. In addition, p auxiliary variables that may be associated with the value of X_i and/or its missingness (δ_i) are also collected and denoted by $\mathbf{Z}_i = \left(Z_i^{(1)}, \dots, Z_i^{(p)}\right)^T$. Then for subject i, the complete data are (D_i, $\mathbf{Z}_i$, δ_i, X_i). When $\delta_i = 1$, the observed data are $O_i = (D_i, \mathbf{Z}_i, \delta_i, X_i)$ and subject i is called a complete case; when $\delta_i = 0$, the observed data are $O_i = (D_i, \mathbf{Z}_i, \delta_i)$ and subject i is called an incomplete case. We denote by $\mathbf{O}$ the collection of observed data for all subjects. When δ_i is independent of X_i conditional on D_i and $\mathbf{Z}_i$, it is a case of MAR; when δ_i is dependent on X_i conditional on D_i and $\mathbf{Z}_i$, it is a case of MNAR.

We are interested in estimating the ROC AUC, which is equivalent to a U-statistic (Bamber, 1975), $\theta = \Pr(X_i > X_j \mid D_i = 1, D_j = 0)$, assuming that the diseased tend to have higher biomarker values. When all data are completely observed, an unbiased estimator of θ is

$$\hat{\theta} = \frac{1}{\sum_{i \neq j} D_i\left(1 - D_j\right)} \sum_{i \neq j} D_i\left(1 - D_j\right) I_{ij},$$

where $I_{ij} = \mathrm{I}(X_i > X_j) + 0.5\mathrm{I}(X_i = X_j)$ with $\mathrm{I}(A)$ equals to 1 if A is true and 0 if A is false. When X is missing for some subjects, a naive extension of the above estimator using only the complete observations (i.e., $\delta_i = 1$) is

$$\hat{\theta}_0 = \frac{1}{\sum_{i \neq j} D_i\left(1 - D_j\right)\delta_i\delta_j} \sum_{i \neq j} D_i\left(1 - D_j\right)\delta_i\delta_j I_{ij}. \tag{1}$$

It is straightforward to verify the following proposition:

Proposition 1: (i) When δ is independent of X given D, $\hat{\theta}_0$ is an unbiased estimator of θ; (ii) when δ is dependent on X given D, then $\hat{\theta}_0$ is subject to potential bias.

We note that (i) includes the case of MCAR and a special case of MAR where δ may depend on D and $\mathbf{Z}$ and is independent of X given D and (ii) includes the case of MNAR and a special case of MAR where δ is dependent on X conditional on D but is independent of X conditional on D and $\mathbf{Z}$. We refer to $\hat{\theta}_0$ as the naive estimator throughout this article.

2.1 Inverse-Weighted Estimator

In the case of MAR, we first study an inverse-weighted estimator,

$$\hat{\theta}_{IW} = \frac{1}{\sum_{i \neq j} \frac{\delta_i\delta_j}{\hat{\pi}_i\hat{\pi}_j} D_i\left(1 - D_j\right)} \sum_{i \neq j} \frac{\delta_i\delta_j}{\hat{\pi}_i\hat{\pi}_j} D_i\left(1 - D_j\right) I_{ij}, \tag{2}$$

where $\hat{\pi}_i$ is an estimate of the probability of observing X_i, namely, $\boldsymbol{\pi}_i = \Pr(\delta_i = 1)$, conditional on $\mathbf{Z}_i$ and D_i under MAR. We denote by $(\mathscr{M}1)$ the working model for $\boldsymbol{\pi}_i$ given $\mathbf{Z}_i$ and D_i with a set of unknown parameters, $\boldsymbol{\alpha}$, and denote by $\mathscr{A}(\boldsymbol{\alpha}; \mathbf{O}) = \sum_i \mathscr{A}_i(\boldsymbol{\alpha}; \mathbf{O})$ the estimating equations for computing the estimate of $\boldsymbol{\alpha}$, namely, $\hat{\boldsymbol{\alpha}}$, based on the observed data. For instance, one can use a logistic regression model for $(\mathscr{M}1)$, i.e. $\mathrm{logit}(\boldsymbol{\pi}_i) = W(\mathbf{Z}_i, D_i; \boldsymbol{\alpha})$ where $W(\mathbf{Z}_i, D_i; \boldsymbol{\alpha})$ is a function of $\mathbf{Z}_i$ and D_i and is parameterized by $\boldsymbol{\alpha}$; $\mathscr{A}(\boldsymbol{\alpha}; \mathbf{O})$ can be taken as the likelihood equations for the logistic regression model. $(\mathscr{M}1)$ is also known as the propensity score model (Rosenbaum and Rubin, 1983). It can be readily shown that if the working model $(\mathscr{M}1)$ is correctly specified, $\hat{\theta}_I$ is a consistent estimator of θ under MAR.

2.2 Doubly Robust Estimators

In the case of MAR, we propose a second estimator

$$\hat{\theta}_{DR} = \frac{1}{\sum_{i \neq j} \frac{\delta_i \delta_j}{\hat{\pi}_i \hat{\pi}_j} D_i(1 - D_j)} \sum_{i \neq j} D_i(1 - D_j) \left\{ \frac{\delta_i \delta_j}{\hat{\pi}_i \hat{\pi}_j} I_{ij} - \frac{\delta_i \delta_j - \hat{\pi}_i \hat{\pi}_j}{\hat{\pi}_i \hat{\pi}_j} E\left(I_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i = 1, D_j = 0\right) \right\} \tag{3}$$

where $\hat{\pi}_i$ is the same as previously defined and $E(I_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i = 1, D_j = 0)$ can be estimated based on the joint conditional distribution of X_i and X_j given the observed data. Specifically, we denote by ($\mathscr{M}2$) the working model for characterizing the *conditional distribution* of X given $\mathbf{Z}$ and D with a set of unknown parameters, $\boldsymbol{\beta}$, and denote by $\mathscr{B}(\boldsymbol{\beta}; \mathbf{O}) = \sum_i \mathscr{B}_i(\boldsymbol{\beta}; \mathbf{O})$ the estimating equations for computing the estimate of $\boldsymbol{\beta}$, namely, $\hat{\boldsymbol{\beta}}$, based on the observed data. We note that the *conditional mean* of X given $\mathbf{Z}$ and D is only part of ($\mathscr{M}2$). It can be readily shown that if either ($\mathscr{M}1$) or ($\mathscr{M}2$) is correctly specified, $\hat{\theta}_{DR}$ is a consistent estimator of θ under MAR.

We consider two options for the working model ($\mathscr{M}2$). In the first option, X given $\mathbf{Z}$ and D is assumed to follow a known parametric distribution with unknown parameters $\boldsymbol{\beta}$. One special case is the Gaussian distribution, i.e., $[X_i \mid \mathbf{Z}_i, D_i] \sim N\left(V(\mathbf{Z}_i, D_i; \boldsymbol{\beta}_1), \sigma_1^2 D_i + \sigma_0^2(1 - D_i)\right)$, where $V(\mathbf{Z}_i, D_i; \boldsymbol{\beta}_1)$ is a function of $\mathbf{Z}_i$ and D_i parameterized by $\boldsymbol{\beta}_1$. Let $\boldsymbol{\beta} = \left(\boldsymbol{\beta}_1^T, \sigma_1^2, \sigma_0^2\right)^T$ denote all parameters of interest, and it follows that

$$X_i - X_j \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i = 1, D_j = 0 \sim N\left(V(\mathbf{Z}_i, D_i = 1; \boldsymbol{\beta}_1) - V\left(\mathbf{Z}_j, D_j = 0; \boldsymbol{\beta}_1\right), \sigma_1^2 + \sigma_0^2\right),$$

and hence

$$E\left\{I_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i = 1, D_j = 0\right\} = \Phi\left(\frac{V(\mathbf{Z}_i, D_i = 1; \boldsymbol{\beta}_1) - V\left(\mathbf{Z}_j, D_j = 0; \boldsymbol{\beta}_1\right)}{\sqrt{\sigma_1^2 + \sigma_0^2}}\right),$$

where $\Phi(\cdot)$ is the cumulative distribution function (c.d.f.) of a standard normal random variable. $\hat{\theta}_{DR}$ can be rewritten as

$$\hat{\theta}_{DR} = \frac{1}{\sum_{i \neq j} \frac{\delta_i \delta_j}{\hat{\pi}_i \hat{\pi}_j} D_i(1 - D_j)} \sum_{i \neq j} D_i(1 - D_j)\left[\frac{\delta_i \delta_j}{\hat{\pi}_i \hat{\pi}_j} I_{ij} - \frac{\delta_i \delta_j - \hat{\pi}_i \hat{\pi}_j}{\hat{\pi}_i \hat{\pi}_j} \Phi\left\{b_{ij}(\hat{\boldsymbol{\beta}})\right\}\right], \tag{4}$$

where $b_{ij}(\hat{\boldsymbol{\beta}}) = \frac{V\left(\mathbf{Z}_i, D_i = 1; \hat{\boldsymbol{\beta}}_1\right) - V\left(\mathbf{Z}_j, D_j = 0; \hat{\boldsymbol{\beta}}_1\right)}{\sqrt{\hat{\sigma}_1^2 + \hat{\sigma}_0^2}}$ and $\hat{\boldsymbol{\beta}}$ can be obtained through, say, linear regression for ($\mathscr{M}2$) using the observed data. From here on, let $\hat{\theta}_{DR}$ denote the doubly robust estimator in Equation (4), which assumes that the conditional distribution of X is Gaussian.

In the second option, suppose $X_i = V(\mathbf{Z}_i, D_i; \boldsymbol{\beta}_1) + \varepsilon_{1i} D_i + \varepsilon_{0i}(1 - D_i)$, where $\{\varepsilon_{1i}, i = 1, \ldots, n_1\}$ and $\{\varepsilon_{0i}, i = 1, \ldots, n_0\}$ are independent and identically distributed (i.i.d.) random errors in the diseased and non-diseased, respectively, and their respective distributions are unknown.

In this case, the conditional distribution of X_i given $\mathbf{Z}_i$ and D_i can be estimated nonparametrically. We denote the set of observed residuals by $\left\{\hat{\varepsilon}_{1k} = X_k - V\left(\mathbf{Z}_k, D_k = 1; \hat{\boldsymbol{\beta}}_1\right), \quad k = 1, \ldots, n_1^o\right\}$ and $\left\{\hat{\varepsilon}_{0l} = X_l - V\left(\mathbf{Z}_l, D_l = 0; \hat{\boldsymbol{\beta}}_1\right), l = 1, \ldots, n_0^o\right\}$ for the diseased and non-diseased, respectively, where n_1^o and n_0^o are the number of subjects with observed X in the diseased and non-diseased, respectively. An empirical sample of the estimated conditional distribution of X_i given $\mathbf{Z}_i$ and D_i can be constructed as $\left\{\widetilde{X}_{ik}^1 = V\left(\mathbf{Z}_i, D_i = 1; \hat{\beta}_1\right) + \hat{\varepsilon}_{1k}, k = 1, \ldots, n_1^o\right\}$ in the diseased and $\left\{\widetilde{X}_{il}^0 = V\left(\mathbf{Z}_i, D_i = 0; \hat{\boldsymbol{\beta}}_1\right) + \hat{\varepsilon}_{0l}, l = 1, \ldots, n_0^o\right\}$ in the non-diseased. $E(I_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i = 1, D_j = 0)$ in Equation (3) can then be estimated using $\frac{1}{n_1^0 n_0^\sigma} \sum_{k=1}^{n_1^0} \sum_{l=1}^{n_0^o} \left\{\mathrm{I}\left(\widetilde{X}_{ik}^1 > \widetilde{X}_{jl}^0\right) + 0.5\mathrm{I}\left(\widetilde{X}_{ik}^1 = \widetilde{X}_{jl}^0\right)\right\}$, where i and j go through all subjects including those with missing X, and we denote the resulting nonparametric estimator of θ by $\hat{\theta}_{DR-N}$. When random errors are not i.i.d., e.g., the variance changes as the mean of X changes, the above procedure needs to be modified accordingly, e.g., performed within strata of the mean of X.

When computing $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$, the weights $\left(\frac{1}{\hat{\pi}_i}\right)$ may be large and unstable, and lead to extra noise in estimation, in particular, when computing the bootstrap SE of $\hat{\theta}_{DR-N}$. Thus, we consider a simple modification to stabilize the weights, namely, replacing $\frac{1}{\hat{\pi}_i}$ with $\frac{1}{\hat{\pi}_i} \frac{n}{\sum_i \delta_i / \hat{\pi}_i}$. When $(\mathscr{M}1)$ is correctly specified, it can be readily shown that $\frac{1}{n} \sum_i \delta_i / \hat{\pi}_i$ converges to 1 in probability, hence $\frac{1}{\hat{\pi}_i} \frac{n}{\sum_i \delta_i / \hat{\pi}_i}$ is equivalent to $\frac{1}{\hat{\pi}_i}$ asymptotically.

2.3 Theoretical Properties

Following our previous notation, we further define the following,

$$\mathscr{U}_{i,j}(\theta, \boldsymbol{\alpha}) \equiv \theta \frac{\delta_i \delta_j}{\pi_i \pi_j} D_i\left(1 - D_j\right) - \frac{\delta_i \delta_j}{\pi_i \pi_j} D_i\left(1 - D_j\right) I_{ij},$$

$$\mathscr{V}_{i,j}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) \equiv \theta \frac{\delta_i \delta_j}{\pi_i \pi_j} D_i\left(1 - D_j\right) - \frac{\delta_i \delta_j}{\pi_i \pi_i} D_i\left(1 - D_j\right) I_{ij} + \frac{\delta_i \delta_j - \pi_i \pi_j}{\pi_i \pi_j} D_i\left(1 - D_j\right) E\left(I_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i, D_j\right),$$

where $\boldsymbol{\pi}_i$ depends on $\boldsymbol{\alpha}$ and $E(I_{ij} \mid \mathbf{Z}_i, \mathbf{Z}_j, D_i, D_j)$ depends on $\boldsymbol{\beta}$. It follows that $\mathscr{U} = \sum_{i \neq j} \mathscr{U}_{i,j}(\theta, \hat{\boldsymbol{\alpha}})$ and $\mathscr{V} = \sum_{i \neq j} \mathscr{V}_{i,j}(\theta, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}})$ are the set of estimating equations for $\hat{\theta}_{IW}$ and $\hat{\theta}_{DR}$, respectively. Let $\boldsymbol{\alpha}_0$ and $\boldsymbol{\beta}_0$ be the probability limits of $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$, respectively, which usually exist.

THEOREM 1: *Under the regularity conditions* (A1)–(A3) *given in* Web Appendix A, *if either or both of* $(\mathscr{M}1)$ and $(\mathscr{M}2)$ *are correctly specified, then* $\sqrt{n}\left(\hat{\theta}_{DR} - \theta\right) \to N(0, \Omega)$ *in distribution, where* $\Omega = Var\left[\left\{E\frac{\delta_i\delta_j}{\pi_i\pi_j}D_i\left(1 - D_j\right)\right\}^{-1} Q_i(\theta, \boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)\right]$, *and*

$$Q_i = -E\left\{\mathscr{V}_{i,j}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) + \mathscr{V}_{j,i}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta}) \mid O_i\right\} + \left[\frac{\partial}{\partial \boldsymbol{\alpha}}E\left\{\mathscr{V}_{i,j}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})\right\}\right] \times \left[\frac{\partial}{\partial \boldsymbol{\alpha}}E\{\mathscr{A}_i(\boldsymbol{\alpha})\}\right]^{-1}\mathscr{A}_i(\boldsymbol{\alpha})$$
$$+ \left[\frac{\partial}{\partial \beta}E\left\{\mathscr{V}_{i,j}(\theta, \boldsymbol{\alpha}, \boldsymbol{\beta})\right\}\right] \times \left[\frac{\partial}{\partial \beta}E\{\mathscr{B}_i(\boldsymbol{\beta})\}\right]^{-1}\mathscr{B}_i(\boldsymbol{\beta}).$$

Ω *can be consistently estimated by* $\hat{\Omega} = \frac{1}{\gamma^2 n}\sum_{i=1}^{n}\hat{Q}_i^2$ *with* $\gamma = \frac{1}{n^2}\sum_{i,j}\frac{\delta_i\delta_j}{\hat{\pi}_i\hat{\pi}_j}D_i\left(1 - D_j\right)$ *and*

$$\hat{Q}_i = -\frac{1}{n}\left[\sum_j\left\{\mathscr{V}_{i,j}\left(\hat{\theta}_{DR}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right) + \mathscr{V}_{j,i}\left(\hat{\theta}_{DR}, \hat{\boldsymbol{\alpha}}, \hat{\boldsymbol{\beta}}\right)\right\}\right] + \frac{1}{n}\left[\sum_{i \neq j}\frac{\partial \mathscr{V}_{i,j}\left(\hat{\theta}_{DR}, \boldsymbol{\alpha}, \hat{\boldsymbol{\beta}}\right)}{\partial \boldsymbol{\alpha}}\Bigg|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}}\right.$$
$$\left.\right]\left[\sum_i\frac{\partial \mathscr{A}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Bigg|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}}\right]^{-1}\mathscr{A}_i(\hat{\boldsymbol{\alpha}}) + \frac{1}{n}\left[\sum_{i \neq j}\frac{\partial \mathscr{V}_{i,j}\left(\hat{\theta}_{DR}, \hat{\boldsymbol{\alpha}}, \boldsymbol{\beta}\right)}{\partial \beta}\Bigg|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}\right]\left[\sum_i\frac{\partial \mathscr{B}_i(\boldsymbol{\beta})}{\partial \beta}\Bigg|_{\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}}\right]^{-1}\mathscr{B}_i(\hat{\boldsymbol{\beta}}).$$

THEOREM 2: *Under the regularity conditions similar to* (A1)–(A3) *given in* Web Appendix A, *if* $(\mathscr{M}1)$ *is correctly specified, then* $\sqrt{n}\left(\hat{\theta}_{IW} - \theta\right) \to N(0, \Omega)$ *in distribution, where* $\Omega = Var\left[\left\{E\frac{\delta_i\delta_j}{\pi_i\pi_j}D_i\left(1 - D_j\right)\right\}^{-1} R_i(\theta, \boldsymbol{\alpha}_0)\right]$, *and*

$$R_i = -E\left\{u_{i,j}(\theta, \boldsymbol{\alpha}) + u_{j,i}(\theta, \boldsymbol{\alpha}) \mid O_i\right\} + \left[\frac{\partial}{\partial \boldsymbol{\alpha}}E\left\{\mathscr{U}_{i,j}(\theta, \boldsymbol{\alpha})\right\}\right] \times \left[\frac{\partial}{\partial \boldsymbol{\alpha}}E\{\mathscr{A}_i(\boldsymbol{\alpha})\}\right]^{-1}\mathscr{A}_i(\boldsymbol{\alpha}).$$

Ω *can be consistently estimated by* $\hat{\Omega} = \frac{1}{\gamma^2 n}\sum_{i=1}^{n}\hat{R}_i^2$ *with* $\gamma = \frac{1}{n^2}\sum_{i,j}\frac{\delta_i\delta_j}{\hat{\pi}_i\hat{\pi}_j}D_i\left(1 - D_j\right)$ *and*

$$\hat{R}_i = -\frac{1}{n}\left[\sum_j\left\{u_{i,j}\left(\hat{\theta}_{IW}, \hat{\boldsymbol{\alpha}}\right) + u_{j,i}\left(\hat{\theta}_{IW}, \hat{\boldsymbol{\alpha}}\right)\right\}\right] + \frac{1}{n}\left[\sum_{i \neq j}\frac{\partial \mathscr{U}_{i,j}\left(\hat{\theta}_{IW}, \boldsymbol{\alpha}\right)}{\partial \boldsymbol{\alpha}}\Bigg|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}}\right]\left[\sum_i\frac{\partial \mathscr{A}_i(\boldsymbol{\alpha})}{\partial \boldsymbol{\alpha}}\Bigg|_{\boldsymbol{\alpha} = \hat{\boldsymbol{\alpha}}}\right]^{-1}\mathscr{A}_i$$
$$(\hat{\boldsymbol{\alpha}}).$$

A sketch of proof for Theorems 1 and 2 is provided in Web Appendix A, which is along the similar lines of Rotnitzky et al. (2006). The underlying idea is to derive the influence functions for $\hat{\theta}_{IW}$ or $\hat{\theta}_{DR}$ by plugging in the influence functions for $\hat{\boldsymbol{\alpha}}$ and $\hat{\boldsymbol{\beta}}$. The consistency of $\hat{\theta}_{DR-N}$ is straightforward to show when either $(\mathscr{M}1)$ or $(\mathscr{M}2)$ holds and its SE can be computed using a bootstrap procedure, which resamples the data with replacement within disease strata.

A few remarks are in order. First, as stated in Proposition 1, $\hat{\theta}_0$ is unbiased when δ is independent of X given D; but if δ and X are associated with $\mathbf{Z}$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ are potentially more efficient when the working models are correctly specified. Second, when δ is dependent on X given D but independent of X given D and $\mathbf{Z}$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ are still consistent under suitable conditions, while $\hat{\theta}_0$ is subject to potential bias. Thirdly, $\hat{\theta}_{DR}$ assumes that the residuals are Gaussian in $(\mathscr{M}2)$ and is subject to model misspecification even if the mean model is correctly specified; $\hat{\theta}_{DR-N}$ does not impose this restriction.

2.4 MNAR: Sensitivity Analysis

We now consider a case of MNAR, where δ is dependent on X conditional on $\mathbf{Z}$ and D; thus, a working model $(\mathscr{M}1)$ that only includes $\mathbf{Z}$ and D is misspecified. We investigate a sensitivity analysis to assess the impact on $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$, as the effect of X on δ is varied. To fix the idea, suppose that logit$(\boldsymbol{\pi}_i) = S(\mathbf{Z}_i, D_i; \boldsymbol{\alpha}_S) + U(X_i, \boldsymbol{\alpha}_X)$, where $\boldsymbol{\alpha}_S$ and $\boldsymbol{\alpha}_X$ are two sets of unknown parameters associated with known functions S and U, respectively. $\boldsymbol{\alpha}_X$ represents the effect of biomarker values on the probability of being missing. Since $\boldsymbol{\alpha}_S$ and $\boldsymbol{\alpha}_X$ can not be jointly estimated using the observed data, we fix $\boldsymbol{\alpha}_X$ at a set of pre-determined values and estimate $\boldsymbol{\alpha}_S$ using the following set of estimating equations,

$$\sum_{i=1}^{n} \left(\frac{\delta_i}{\pi_i} - 1\right) W(\mathbf{Z}_i, D_i), \tag{5}$$

where $W(\mathbf{Z}_i, D_i)$ is an arbitrary known vector function with the same dimension as $\boldsymbol{\alpha}_S$. For instance, if $S(\mathbf{Z}_i, D_i; \boldsymbol{\alpha}_S) = \boldsymbol{\alpha}_S W(\mathbf{Z}_i, D_i)$, then $W(\mathbf{Z}_i, D_i)$ is the covariate vector for i, which may include interaction terms. Compared to the likelihood equations for the logistic regression, one advantage of the estimation equations (5) is that $\boldsymbol{\pi}_i$ is not needed when X_i is missing. For every pre-determined value of $\boldsymbol{\alpha}_X$, we can compute $\hat{\boldsymbol{\alpha}}_S$ using (5) and $\hat{\pi}_i$ for subjects with observed X_i; subsequently we can compute $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$, all of which do not need $\hat{\pi}_i$ for subjects with missing X_i. This procedure is repeated for a grid of $\boldsymbol{\alpha}_X$ values, and the resulting estimators are compared to assess the impact of $\boldsymbol{\alpha}_X$ and hence the impact of MNAR. $U(X_i, \boldsymbol{\alpha}_X) = 0$ corresponds to the case of MAR, and $U(X_i, \boldsymbol{\alpha}_X) \neq 0$ corresponds to the case of MNAR. In this sensitivity analysis, we do not assume that the estimation of the parameters of $(\mathscr{M}2)$ is not affected by MNAR. To simplify the sensitivity analysis and, in particular, avoid performing sensitivity analysis for two working models, we exploit the doubly robust property, i.e., if $(\mathscr{M}1)$ is correctly specified then the proposed estimators are consistent, and focus on $(\mathscr{M}1)$ only.

3. Simulation studies

We conducted simulations to evaluate the finite sample performance of the proposed estimators, first in the case of MAR where δ is independent of X given D and $\mathbf{Z}$, then in the case of MNAR where δ is dependent on X given D and $\mathbf{Z}$. In our simulations, $\hat{\theta}_0$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ were compared. In addition, we considered another estimator, namely, $\hat{\theta}_{IMP} = \frac{1}{\sum_{i \neq j} D_i(1-D_j)} \sum_{i \neq j} D_i(1-D_j)\left[\delta_i\delta_j I_{ij} - (\delta_i\delta_j - 1)\Phi\left\{b_{ij}(\hat{\boldsymbol{\beta}})\right\}\right]$, which only relies on $(\mathscr{M}2)$ and is not doubly robust. While it is not of primary interest in this article, $\hat{\theta}_{IMP}$ under the correctly specified $(\mathscr{M}2)$ can be used as an optimal benchmark for efficiency as suggested by a referee. To benchmark bias and loss of efficiency due to missing data, a so-called gold standard (GS) estimator was also computed, i.e., $\hat{\theta}_{GS} = \frac{1}{\sum_{i \neq j} D_i(1 - D_j)} \sum_{i=1}^{n} \sum_{j=1}^{n} D_i(1 - D_j) I_{ij}$, which uses the underlying "true" biomarker values for all subjects and is not applicable in real data analysis. In Tables 1–3, modified weights as described in Section 2.2 were used for $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$; to compute the standard error for $\hat{\theta}_{DR-N}$, we used 200 bootstrap datasets randomly sampled with replacement from the data while stratified on the disease status. In each simulation, we generated a random sample of $n = 200$ independent subjects with an equal number of diseased and non-diseased subjects. For each simulation setting, 500 Monte Carlo datasets were generated and the results were summarized using the following measures: the mean relative bias (RB), mean of the standard error estimates (SE), Monte Carlo standard deviation of parameter estimates (SD), square root of mean squared errors (SMSE) and coverage rate (CR) of 95% Wald's confidence interval using a logistic transform of θ as suggested in Pepe (2003) (Ch. 5).

3.1 MAR: δ independent of X given D and Z

Under MAR, we considered two settings, namely, δ dependent on X given D and δ independent of X given D. Corresponding to each setting, we generated the auxiliary variables, $\mathbf{Z}_1 = \left(Z_1^{(1)}, Z_1^{(2)}, Z_1^{(3)}\right)$, which are associated with δ, and $\mathbf{Z}_2 = \left(Z_2^{(1)}, Z_2^{(2)}, Z_2^{(3)}\right)$, which are associated with X. In the first setting, $\mathbf{Z}_1 = \mathbf{Z}_2$ and they were generated from a multivariate Gaussian distribution with a mean $\mu_Z = (3, -2, -1)$ and a variance matrix $\Sigma_Z =$ diag(0.25, 0.25, 0.25), which implies that δ is dependent on X given D and hence $\hat{\theta}_0$ is subject to potential bias. In the second setting, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ were generated from two independent multivariate Gaussian distributions with the same mean and variance as in the first setting, which implies that δ is independent of X given D and hence $\hat{\theta}_0$ is unbiased. Next, we generated X as follows, $X = \beta_0 + \beta_1 D + \boldsymbol{\beta}_2 \mathbf{Z}_2 + \boldsymbol{\beta}_3 D\mathbf{Z}_2 + \varepsilon$ with $\beta_0 = 1$, $\beta_1 = 2.5$, $\boldsymbol{\beta}_2 = (3, 3, 3)$, and $\boldsymbol{\beta}_3 = (.5, .5, .5)$, which is the true underlying model for $(\mathscr{M}2)$. Two different residual distributions were considered so that we could compare the performance of $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$; specifically, $\varepsilon \sim N(0, \sigma^2)$ or $\varepsilon = 20\{\eta - E(\eta)\}$ with $\eta \sim$ Beta(5, 1). The resulting true θ is 0.722 for Gaussian ε and 0.675 for non-Gaussian ε. Subsequently, we generated the missing indicator δ from a Bernoulli distribution with mean $\boldsymbol{\pi}$ which satisfies logit$(\boldsymbol{\pi}) = \alpha_0 + \alpha_1 D + \boldsymbol{\alpha}_2 \mathbf{Z}_1 + \boldsymbol{\alpha}_3 D\mathbf{Z}_1$ with $\alpha_0 = 0.3$, $\alpha_1 = 0.3$, $\boldsymbol{\alpha}_2 = (0.4, 0.5, 0.3)$, and $\boldsymbol{\alpha}_3 = (-0.7, -0.7, -0.9)$; this is the underlying true model for $(\mathscr{M}1)$. The resulting average probability of missing X is 66.4% in the diseased group and 55.8% in the non-diseased group.

When computing $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$, we fitted the two working models for δ and X, namely, $(\mathscr{M}1)$ and $(\mathscr{M}2)$, under the following four scenarios: 1) the mean structure is correctly specified for both working models, i.e., $\mathbf{Z}_1$ and D are included in $(\mathscr{M}1)$, and $\mathbf{Z}_2$ and D are included in $(\mathscr{M}2)$; 2) the mean structure is misspecified for $(\mathscr{M}1)$, i.e., only $Z_1^{(1)}$ and D are used in $(\mathscr{M}1)$; 3) the mean structure is misspecified for $(\mathscr{M}2)$, i.e., only $Z_2^{(1)}$ and D are used in $(\mathscr{M}2)$; and 4) the mean structure is misspecified for both working models, i.e., only

$Z_1^{(1)}$ and D are included in ($\mathscr{M}1$) and only $Z_2^{(1)}$ and D are included in ($\mathscr{M}2$). We note that $\hat{\theta}_{DR}$ assumes that X follows Gaussian distributions. Consequently, if the residuals for X follow a Gaussian distribution, e.g. $\varepsilon \sim \mathrm{N}(0, \sigma^2)$, then the correct specification of the mean structure in ($\mathscr{M}2$) also indicates the correct specification of the conditional distribution for X when computing $\hat{\theta}_{DR}$. However, if the residual distribution is not Gaussian, e.g., $\varepsilon = 20\{\eta - E(\eta)\}$ with $\eta \sim \mathrm{Beta}(5, 1)$, the conditional distribution for X is misspecified when computing $\hat{\theta}_{DR}$, even if the mean structure is correctly specified in ($\mathscr{M}2$). Since $\hat{\theta}_{DR-N}$ is robust to the mis-specification of distributions of the residuals for X, it should remain consistent in both cases.

3.1.1 The case of δ dependent on X given D.

—In this setting, we let $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be identical, hence δ is dependent on X given D. Table 1 presents the results for two different residual distributions for X. We first discuss the case of Gaussian ε. $\hat{\theta}_0$ shows a large RB of 11.6% with a low coverage rate of 70.0%. $\hat{\theta}_{IW}$ exhibits negligible bias and a CR close to the nominal level when ($\mathscr{M}1$) is correctly specified; however, its bias becomes substantial and CR degrades considerably to 78.6% when ($\mathscr{M}1$) is misspecified. When at least one working model is correctly specified, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ show negligible bias that is comparable to $\hat{\theta}_{GS}$ and good coverage properties. In particular, as long as ($\mathscr{M}2$) is correctly specified, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ are more efficient than $\hat{\theta}_{IW}$, and is almost as efficient as $\hat{\theta}_{IMP}$; in this case, negligible loss of efficiency is observed even if ($\mathscr{M}1$) is misspecified. By contrast, when ($\mathscr{M}2$) is misspecified and ($\mathscr{M}1$) is correctly specified, the loss of efficiency is considerable for $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$. These observations are consistent with what have been reported in the literature, i.e., the correct specification of ($\mathscr{M}2$) for X is more important in terms of improving efficiency of $\hat{\theta}_{DR}$. When both working models are misspecified, the bias and MSE of $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ are still similar to or less than those of $\hat{\theta}_{IW}$ or $\hat{\theta}_0$.

When the residuals are not Gaussian, ($\mathscr{M}2$) is always misspecified for $\hat{\theta}_{DR}$. Our results in Table 1 show that $\hat{\theta}_{DR}$ is fairly robust to the misspecified distribution of ε as long as the conditional mean of X in ($\mathscr{M}2$) is correctly specified. In addition, most observations for Gaussian ε are still true for non-Gaussian ε. In this case, $\hat{\theta}_{IMP}$ serves as an approximate benchmark for efficiency, since $\hat{\theta}_{IMP}$ is also fairly robust to a mis-specified distribution for ε and it is generally difficult to obtain an exact “imputation” estimator when ε is non-Gaussian. Similar results were observed in our additional simulations with other non-Gaussian distributions for ε, say, χ^2 distribution.

3.1.2 The case of δ independent of X given D.

—In this setting, $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are two separate sets of auxiliary variables, hence δ is independent of X given D. Table 2 presents the results for both Gaussian and non-Gaussian residuals. In all cases, all estimators exhibit negligible bias and satisfactory coverage properties, which is consistent with our discussion in Section 2. Again, as long as ($\mathscr{M}2$) is (approximately) correctly specified, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ are almost as efficient as $\hat{\theta}_{IMP}$; they perform no worse than $\hat{\theta}_{IW}$ and $\hat{\theta}_0$ in other settings. As with the case of δ dependent on X given D in Section 3.1.1, the results are very similar for two different types of residual distributions for X.

We repeated the simulations in Tables 1 and 2 using the original weights (Web Appendix B), and the results are almost the same except that the performance of the bootstrap SE for $\hat{\theta}_{DR-N}$ deteriorates somewhat.

3.2 MNAR: δ dependent on X given D and Z

We now consider the case of MNAR where δ is dependent on X conditional on D and $\mathbf{Z}$, i.e., the true model for δ is $\mathrm{logit}(\pi) = \alpha_0 + \alpha_Z Z_1^{(3)} + \alpha_D D + \alpha_X X$ with $(\alpha_0, \alpha_Z, \alpha_D, \alpha_X) = (-1, 0.2, 0.5, 0.3)$. The rest of the simulation setup is identical to that in Section 3.1. The resulting average probability of missing X is 57.4% in the diseased group and 31.5% in the non-diseased group. We focused on the case where $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are identical and ε is Gaussian; in this case, the true θ remains 0.722. Our primary goal is to compare $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ with their corresponding sensitivity estimators as described in Section 2.4, namely, $\hat{\theta}_{IW-S}$, $\hat{\theta}_{DR-S}$ and $\hat{\theta}_{DR-N-S}$, for which the estimating equations (5) were used to estimate $\boldsymbol{\alpha}_S = (\alpha_0, \alpha_Z, \alpha_D)$ with α_X fixed at its true value. The rest of estimating procedures remain the same for all estimators. As with the case of MAR in Section 3.1, we investigated the impact of the mis-specified $(\mathscr{M}1)$ and/or $(\mathscr{M}2)$; specifically, we considered a misspecified $(\mathscr{M}1)$ that includes $Z_1^{(1)}$ and D and a misspecified $(\mathscr{M}2)$ that includes only $Z_2^{(3)}$ and D. We also note that X is included as a covariate in $(\mathscr{M}1)$ for $\hat{\theta}_{IW-S}$, $\hat{\theta}_{DR-S}$ and $\hat{\theta}_{DR-N-S}$, but not for $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ or $\hat{\theta}_{DR-N}$. Thus, when D and the correct subset of $\mathbf{Z}_1$ (i.e., $Z_1^{(3)}$) are included in $(\mathscr{M}1)$, $(\mathscr{M}1)$ is correctly specified for $\hat{\theta}_{IW-S}$, $\hat{\theta}_{DR-S}$ and $\hat{\theta}_{DR-N-S}$, but is misspecified for $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$.

Table 3 presents the simulation results. First, $\hat{\theta}_0$ again exhibits substantial bias under MNAR. We now compare $\hat{\theta}_{IW}$ and $\hat{\theta}_{IW-S}$. When $(\mathscr{M}1)$ does not account for the effect of X, $\hat{\theta}_{IW}$ shows considerable bias even if $(\mathscr{M}1)$ includes D and the correct subset of $\mathbf{Z}_1$. On the other hand, $\hat{\theta}_{IW-S}$, which accounts for the effect of X, shows negligible bias. Next, we compare $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ with $\hat{\theta}_{DR-S}$ and $\hat{\theta}_{DR-N-S}$. When correct subsets of $\mathbf{Z}$ and D are included in both working models, $(\mathscr{M}1)$ is still misspecified for $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$. However, since $(\mathscr{M}2)$ is correctly specified, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ exhibit negligible bias and good coverage properties as a result of their double robustness, and their efficiency is comparable to that of $\hat{\theta}_{DR-S}$ and $\hat{\theta}_{DR-N-S}$. These results still hold when $(\mathscr{M}1)$ includes the incorrect subset of auxiliary variables and $(\mathscr{M}2)$ is correctly specified. When an incorrect subset of $\mathbf{Z}_2$ is included in $(\mathscr{M}2)$, both working models are misspecified for $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$; consequently, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ exhibit considerable bias. In all three settings, $\hat{\theta}_{DR-S}$ and $\hat{\theta}_{DR-N-S}$ show negligible bias, but their SDs increase when $(\mathscr{M}2)$ is misspecified, which is consistent with the earlier findings that $(\mathscr{M}2)$ is more important in terms of improving efficiency.

4. Data Analysis

We illustrate our methods using an observational psychiatric study, which was concerned with the impact of maternal depression during pregnancy on infant outcomes. In this study, participants were enrolled no later than week 28 of gestation and evaluated at each trimester across pregnancy. As part of the study, the presence (or absence) of a major depressive episode (disease status, D) was determined at each visit by the Mood Module of the Structured Clinical Interview for DSMIV Axis I Disorders (SCID) (First et al., 2002), which needs to be administered by a trained research professional and is considerably more time-consuming and difficult to obtain in practice. At the same time, some subjects also completed the self-rated Edinburgh Postnatal Depression Scale (EPDS) (Cox and Holden, 1987) at each visit.

In female mental health research, several rating scales have been developed for identifying postpartum depression (Fergerson et al., 2002; Perfetti et al., 2004), and in particular the self-rated EPDS has emerged as a widely-used instrument for postpartum depression screening and detection (Austin et al., 2005; Felice et al., 2006), which can be obtained fairly easily in practice. In contrast, there are no validated tools to assess depression during pregnancy. In practice, the EPDS, developed for postpartum use, has been increasingly used to identify depression during pregnancy and to screen for those at risk for developing depression during pregnancy. While not designed for such purpose, data collected from this study have been recently used to evaluate EPDS as a biomarker for the diagnosis of maternal depression throughout pregnancy. For the purpose of illustration, we focus on the data collected from the second trimester; a subset of the study population who had data in the second trimester was used and the sample size is $n = 517$ in the analysis. The outcome of interest is the presence of a major depressive episode (D) and is confirmed for all subjects, whereas EPDS is the biomarker of interest and is missing in 79% of the subjects. Additional auxiliary variables were also measured in this study including the mother's age, race, marital status and eduction level, whether or not it was the first pregnancy. In addition, a research interviewer masked to treatment status administered the Structured Interview Guide for the Hamilton Rating Scale for Depression to obtain 17-item (HRSD17), which is known to be highly correlated with EPDS. These variables are treated as auxiliary variables ($\mathbf{Z}$) and are used to build $(\mathscr{M}1)$ and $(\mathscr{M}2)$.

We conducted a sensitivity analysis for $\hat{\theta}_0$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ as described in Section 2.4. Specifically, we considered a $(\mathscr{M}1)$ that is similar to what is discussed in Section 2.4, i.e., $\mathrm{logit}(\pi) = \alpha_S^T W(\mathbf{Z}, D) + \alpha_X X$, where $W(\mathbf{Z}, D)$ include the intercept and interaction terms between auxiliary variables $\mathbf{Z}$ and D. In fitting this $(\mathscr{M}1)$, estimating equations (5) were used with α_X fixed at −1, 0 and 1, where $\alpha_X = 0$ corresponds to the case of MAR, and $\alpha_X = -1$ or 1 correspond to the case of MNAR. In our analysis, all continuous variables including X were standardized to have mean 0 and unit standard deviation. Consequently, α_X captures the effect of a one-SD change in X. Table 4 presents the results using modified weights. The impact of different α_X values is moderate on $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$, and $\hat{\theta}_{DR-N}$, and the estimates using different methods including $\hat{\theta}_0$ are comparable. It indicates that the missingness of X

(δ) is likely close to be independent of X given D. Nevertheless, $\hat{\theta}_{DR}$, which incorporates information from auxiliary variables, is more efficient than the other estimators. Since the proportion of missing data is very high in the data, the bootstrap SE of $\hat{\theta}_{DR-N}$ is greater than the SE of $\hat{\theta}_{DR}$, but it is still smaller than the SE of $\hat{\theta}_{IW}$. We repeated this analysis using the original weights in Web Appendix C; while the main results remain similar, a larger bootstrap SE for $\hat{\theta}_{DR-N}$ is observed as a result of large and unstable weights.

With $\hat{\theta}_{DR}$ ranging from 0.841 to 0.873, our results suggest that EPDS has very good discriminative power during the second trimester. However, in this study, only a subset of the study population had depression status confirmed during each perinatal window. As a result, in addition to missing values in the biomarker, the verification bias is potentially in play as well. Furthermore, both the rating scale and the presence of a major depressive episode were repeatedly measured through the pregnancy. Therefore, it is of substantial interest in the future studies to investigate methods that can account for both missing biomarker values as well as verification bias and accommodate repeatedly measured biomarker values and disease status when estimating the ROC AUC.

5. Discussion

We have proposed and contrasted several estimators of the ROC AUC when the biomarker value is missing for some subjects. Our numerical studies show that the doubly robust estimators perform as well as or better than other estimators in all cases even when both working models are misspecified. $\hat{\theta}_{DR}$ is also fairly robust to the misspecified residual distribution for the biomarker variable (X). Since only ranks of X are used in estimating θ, the correct specified conditional mean is more important and the impact of a misspecified residual distribution may be limited given the correctly specified conditional mean. The bootstrap procedure for obtaining SE of $\hat{\theta}_{DR-N}$ is computationally more expensive and also makes it more susceptible to large and unstable weights. Thus, in practice, we recommend the use of $\hat{\theta}_{DR}$ and stabilized weights such as ours, and emphasize the importance of identifying (approximately) correct ($\mathscr{M}2$). We also note that $\hat{\theta}_{DR}$ can readily accommodate categorical biomarker values, e.g., a baseline logit model (Agresti, 2002) can be used to model the conditional distribution of a categorical biomarker variable.

More recently, Cao et al. (2009) investigated alternative doubly robust estimators for estimating a population mean; their methods achieve minimum variance under incorrectly specified ($\mathscr{M}2$) and correctly specified ($\mathscr{M}1$), and they do not suffer from large and unstable weights. While their enhanced model for ($\mathscr{M}1$) can be readily adopted in our methods as an alternative to alleviate the problem of large and unstable weights, it is more involved to extend their approach of minimizing variance under misspecified ($\mathscr{M}2$) and correctly specified ($\mathscr{M}1$) to the estimation of the ROC AUC as complications arise from the use of U-statistic in our methods. Potential future research may also include extending sensitivity analysis to ($\mathscr{M}2$) and investigating more complicated missing patterns, e.g., auxiliary variables are also missing and missingness is not monotone, for which an imputation approach may be more practical.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgements

We thank Editor Verbeke, an associate editor and two referees for their insightful suggestions which greatly improved an earlier draft of this manuscript.

References

Agresti A (2002). Categorial Data Analysis, 2nd Edition. John Wiley & Sons.

Austin M, D. H-P, Saint K, and Parker G (2005). Antenatal screening for the prediction of postnatal depression: validation of a psychosocial pregnancy risk questionnaire. Acta Psychiatr Scand. 112, 310–317. [PubMed: 16156839]

Bamber D (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. Journal of Mathematical Psychology. 12, 387C415.

Cao W, Tsiatis A, and Davidian M (2009). Improving efficiency and robustness of the doubly robust estimator for a population mean with incomplete data. Biometrika 96, 723–734. [PubMed: 20161511]

Cox J and Holden J (1987). Detection of postnatal depression. development of the 10-item edinburgh postnatal depression scale. Br J Psychiatry. 150, 782–786. [PubMed: 3651732]

Felice E, Saliba J, Grech V, and Cox J (2006). Validation of the maltese version of the edinburgh postnatal depression scale. Arch Womens Ment Health. 9, 75–80. [PubMed: 16172837]

Fergerson S, Jamieson D, and Lindsay M (2002). Diagnosing postpartum depression: can we do better? Am J Obstet Gynecol. 186, 899–902. [PubMed: 12015507]

First M, Spitzer R, Gibbon M, and Williams J (2002). Structured Clinical Interview for DSM-IV-TR Axis I Disorders, Patient Edition (SCID-IP, 11/2002 Revision). Washington, DC: American Psychiatric Press.

Fluss R, Reiser B, Faraggi D, and Rotnitzky A (2009). Estimation of the roc curve under verification bias. Bio-1metrical Journal. 51(3), 475–490.

Green D and Swets J (1966). Signal detection theory and psychopysics. Wiley, New York.

Kosinski A and Barnhart H (2003). Accounting for non-ignorable verification bias in assessment of diagnostic test. Biometrics 59, 163–171. [PubMed: 12762453]

Little R and Rubin D (2002). Statistical Analysis with Missing Data. 2nd Edition. Wiley &. Sons.

Pepe M (2003). The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford: University Press.

Perfetti J, Clark R, and Fillmore C (2004). Postpartum depression: identification, screening, and treatment. Wis Med J. 103, 56–63.

Robins J, Rotnitzky A, and Zhao L (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association. 89, 846–866.

Rosenbaum P and Rubin D (1983). The central role of the propensity score in observational studies for causal effects. Biometrika 70, 41–55.

Rotnitzky A, Faraggi D, and Schisterman E (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. Journal of the American Statistical Association. 101, 1276–1288.

Rotnitzky A and Robins J (1997). Analysis of semiparametric regression models with non-ignorable nonresponse. Statistics in Medicine. 16, 81–102. [PubMed: 9004385]

Scharfstein D, Rotnitzky A, and Robins J (1999). Adjusting for nonignorable dropout using semiparametric nonresponse models (with discussion). Journal of the American Statistical Association 94, 1096–1120.

Zhou X (1993). Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communication in Statistics-Theory and Methods. 22, 3177–3198.

Zhou X (1994). Effect of verification bias on positive and negative predictive values. Statistics in Medicine. 13, 1737–1745. [PubMed: 7997707]

Zhou X (1998). Correcting for verification bias in studies of a diagnostic test's accuracy. Statistical Methods in Medical Research. 7, 337–353. [PubMed: 9871951]

Zweig M and Campbell G (1993). Receiver-operating characteristic (roc) plots: A fundamental evaluation tool in clinical medicine. Clinical Chemistry. 39, 561–577. [PubMed: 8472349]

Table 1

Results of simulation study under MAR: comparison of $\hat{\theta}_0$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ using modified weights, when $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are identical. ε is Gaussian (i.e., $\varepsilon \sim N(0, 1)$) or non-Gaussian (i.e., $\varepsilon = 20\{\eta - E(\eta)\}$ with $\eta \sim$ Beta(5, 1)). True θ is 0.722 for Gaussian ε and 0.675 for non-Gaussian ε. The details of true models and misspecified working models are provided in Section 3.1.

	Gaussian ε					**non-Gaussian ε**				
	RB (%)	**SE**	**SD**	**SMSE**	**CR (%)**	**RB (%)**	**SE**	**SD**	**SMSE**	**CR (%)**
$\hat{\theta}_{GS}$	−0.2	0.036	0.037	0.037	94.0	0.0	0.038	0.038	0.038	95.8
$\hat{\theta}_0$	11.6	0.050	0.054	0.099	70.0	10.8	0.057	0.056	0.092	80.4
Both mean models correctly specified										
$\hat{\theta}_{IMP}$	−0.1	0.040	0.042	0.042	95.0	−0.8	0.054	0.054	0.054	94.8
$\hat{\theta}_{IW}$	0.5	0.048	0.052	0.052	93.0	1.0	0.056	0.058	0.059	95.0
$\hat{\theta}_{DR}$	0.0	0.045	0.043	0.043	96.4	0.5	0.057	0.055	0.055	96.4
$\hat{\theta}_{DR-N}$	0.0	0.042	0.043	0.043	94.4	0.5	0.056	0.056	0.056	96.0
Mean model for $(\mathscr{M}1)$ misspecified										
$\hat{\theta}_{IW}$	8.4	0.050	0.054	0.081	78.6	8.0	0.056	0.058	0.079	84.8
$\hat{\theta}_{DR}$	0.0	0.040	0.043	0.043	94.0	0.6	0.052	0.055	0.055	94.4
$\hat{\theta}_{DR-N}$	0.0	0.041	0.043	0.043	95.4	0.4	0.056	0.055	0.055	95.8
Mean model for $(\mathscr{M}2)$ misspecified										
$\hat{\theta}_{DR}$	0.5	0.054	0.050	0.050	96.2	0.9	0.061	0.058	0.058	96.2
$\hat{\theta}_{DR-N}$	0.5	0.050	0.050	0.050	94.0	0.9	0.059	0.058	0.058	95.6
Both mean models misspecified										
$\hat{\theta}_{DR}$	8.4	0.050	0.053	0.081	78.6	7.9	0.057	0.058	0.079	86.0
$\hat{\theta}_{DR-N}$	8.4	0.051	0.053	0.081	79.0	7.9	0.058	0.058	0.079	86.2

Table 2

Results of simulation study under MAR: comparison of $\hat{\theta}_0$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$ and $\hat{\theta}_{DR-N}$ using modified weights, when $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are independent. ε is Gaussian (i.e., $\varepsilon \sim N(0, 1)$) or non-Gaussian (i.e., $\varepsilon = 20\{\eta \sim E(\eta)\}$ with $\eta \sim \text{Beta}(5, 1)$). True θ is 0.722 for Gaussian ε and 0.675 for non-Gaussian ε. The details of true models and misspecified working models are provided in Section 3.1.

	Gaussian ε					**non-Gaussian ε**				
	RB (%)	**SE**	**SD**	**SMSE**	**CR (%)**	**RB (%)**	**SE**	**SD**	**SMSE**	**CR (%)**
$\hat{\theta}_{GS}$	0.2	0.036	0.035	0.035	96.0	0.2	0.038	0.036	0.036	96.0
$\hat{\theta}_0$	−0.1	0.059	0.057	0.057	95.8	0.4	0.063	0.061	0.061	96.4
Both mean models correctly specified										
$\hat{\theta}_{IMP}$	0.1	0.039	0.040	0.040	94.2	−0.6	0.051	0.050	0.050	95.8
$\hat{\theta}_{IW}$	−0.1	0.059	0.060	0.059	94.8	0.3	0.062	0.065	0.065	94.6
$\hat{\theta}_{DR}$	0.2	0.044	0.040	0.040	96.4	0.1	0.056	0.053	0.053	95.8
$\hat{\theta}_{DR-N}$	0.2	0.041	0.041	0.041	95.4	0.1	0.056	0.053	0.053	95.2
Mean model for $(\mathscr{M}1)$ misspecified										
$\hat{\theta}_{IW}$	−0.2	0.058	0.059	0.059	95.0	0.3	0.062	0.062	0.062	95.2
$\hat{\theta}_{DR}$	0.2	0.041	0.040	0.040	95.0	0.2	0.053	0.051	0.051	95.8
$\hat{\theta}_{DR-N}$	0.2	0.041	0.040	0.040	94.8	0.1	0.054	0.051	0.051	96.2
Mean model for $(\mathscr{M}2)$ misspecified										
$\hat{\theta}_{DR}$	−0.2	0.059	0.055	0.055	96.0	0.4	0.063	0.062	0.062	95.4
$\hat{\theta}_{DR-N}$	−0.2	0.057	0.055	0.055	96.4	0.4	0.063	0.062	0.062	95.2
th Mean Models misspecified										
$\hat{\theta}_{DR\cdot}$	−0.2	0.054	0.054	0.054	95.0	0.4	0.059	0.059	0.059	95.2
$\hat{\theta}_{DR-N}$	−0.2	0.055	0.054	0.054	95.8	0.4	0.061	0.059	0.059	95.8

Table 3

Results of simulation study under MNAR: comparison of $\hat{\theta}_0$, $\hat{\theta}_{IW}$, $\hat{\theta}_{DR}$, $\hat{\theta}_{DR-N}$, $\hat{\theta}_{IW-S}$, $\hat{\theta}_{DR-S}$, and $\hat{\theta}_{DR-N-S}$ using modified weights, when $\varepsilon \sim N(0, 1)$ and $\mathbf{Z}_1$ and $\mathbf{Z}_2$ are identical. True θ is 0.722. The details of true models and misspecified working models are provided in Section 3.2.

	RB (%)	SE	SD	SMSE	CR (%)
$\hat{\theta}_{GS}$	−0.2	0.036	0.037	0.037	94.0
$\hat{\theta}_0$	−8.1	0.053	0.055	0.080	78.2
Correct subset of $\mathbf{Z}_1$ and D included in $(\mathscr{M}1)$					
$\hat{\theta}_{IW}$	−5.8	0.052	0.055	0.069	85.4
$\hat{\theta}_{IW-S}$	−1.0	0.049	0.052	0.053	93.0
Correct subset of $\mathbf{Z}$ and D included in both models					
$\hat{\theta}_{DR}$	−0.5	0.039	0.040	0.041	95.0
$\hat{\theta}_{DR-N}$	−0.5	0.039	0.040	0.040	94.4
$\hat{\theta}_{DR-S}$	−0.2	0.038	0.039	0.039	93.8
$\hat{\theta}_{DR-N-S}$	−0.2	0.038	0.039	0.039	93.6
Incorrect subset of $\mathbf{Z}_1$ and D included in $(\mathscr{M}1)$					
$\hat{\theta}_{DR}$	−0.6	0.039	0.040	0.041	94.6
$\hat{\theta}_{DR-N}$	−0.5	0.039	0.040	0.040	93.2
$\hat{\theta}_{DR-S}$	−0.2	0.038	0.039	0.039	94.0
$\hat{\theta}_{DR-N-S}$	−0.2	0.038	0.039	0.039	93.4
Incorrect subset of $\mathbf{Z}_2$ and D included in $(\mathscr{M}2)$					
$\hat{\theta}_{DR}$	−5.3	0.049	0.053	0.065	85.0
$\hat{\theta}_{DR-N}$	−5.3	0.050	0.053	0.065	86.2
$\hat{\theta}_{DR-S}$	−0.8	0.047	0.049	0.049	92.8
$\hat{\theta}_{DR-N-S}$	−0.8	0.048	0.049	0.049	93.8

Table 4

Sensitivity analysis using the modified weights for estimating the ROC AUC (θ) in the psychiatric study

	Estimate	**SE**
$\hat{\theta}_0$	**0.861**	**0.038**

	$\alpha_X = -1$		$\alpha_X = 0$		$\alpha_X = 1$	
	Estimate	SE	Estimate	SE	Estimate	SE
$\hat{\theta}_{IW}$	0.864	0.037	0.851	0.040	0.849	0.042
$\hat{\theta}_{DR}$	0.873	0.028	0.852	0.030	0.841	0.032
$\hat{\theta}_{DR-N}$	0.873	0.035	0.852	0.038	0.841	0.038