Published in final edited form as:

Biometrics. 2010 December ; 66(4): 1119–1128. doi:10.1111/j.1541-0420.2010.01397.x.

A model for adjusting for nonignorable verification bias in estimation of ROC curve and its area with likelihood-based approach

Danping Liu* and

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

Xiao-Hua Zhou*

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

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

Summary

In estimation of the ROC curve, when the true disease status is subject to nonignorable missingness, the observed likelihood involves the missing mechanism given by a selection model. In this paper, we proposed a likelihood-based approach to estimate the ROC curve and the area under ROC curve when the verification bias is nonignorable. We specified a parametric disease model in order to make the nonignorable selection model identifiable. With the estimated verification and disease probabilities, we constructed four types of empirical estimates of the ROC curve and its area based on imputation and reweighting methods. In practice, a reasonably large sample size is required to estimate the nonignorable selection model in our settings. Simulation studies showed that all the four estimators of ROC area performed well, and imputation estimators were generally more efficient than the other estimators proposed. We applied the proposed method to a data set from research in the Alzheimer's disease.

Keywords

Alzheimer's disease; nonignorable missing data; ROC curve; verification bias

1. Introduction

The receiver operating characteristic (ROC) curve is often used to evaluate the accuracy of a medical diagnostic test. In estimation of an ROC curve, it is often the case that the gold standard or true disease status is not verified for every patient. Using only the verified patients to estimate the ROC curve would result in verification bias (Begg and Greenes, 1983). A common assumption in adjusting for verification bias is that selection for disease verification does not depend on disease status, given test results and other observed covariates, namely, missing at random (MAR) assumption. Published methods include but are not limited to: Begg and Greenes (1983), Begg (1987), Zhou (1996), Zhou (1998), Rodenber and Zhou (2000), Alonzo and Pepe (2005), and He, Lyness and McDermott (2009).

*dpliu@u.washington.edu

*azhou@u.washington.edu

Supplementary Material

Web Appendices referenced in Sections 1, 3, 4 and 5 are available under the Paper Information link at the *Biometrics* website <http://www.biometrics.tibs.org>.

However, disease verification may also depend on some unobserved covariates related to disease, and hence the MAR assumption may not hold, which is known as the nonignorable missing data mechanism. In this situation, the observed likelihood involves the missing data mechanism given by a selection model. Baker (1995) first proposed the selection model under nonignorable verification bias and derived estimators of sensitivity and specificity for multiple binary tests. Zhou and Rodenberg (1998) specified a dependence structure between verification and disease status and obtained the maximum likelihood estimate of the ROC curve. Kosinski and Barnhart (2003) incorporated covariates in the selection model and estimated the sensitivity and specificity parameters using an EM algorithm. Zhou and Castelluccio (2003 and 2004) proposed a profile likelihood approach combined with an EM algorithm to obtain a nonparametric estimator of the ROC curve.

Recently, Rotnitzky, Faraggi, and Schisterman (2006) developed a doubly robust estimator of the area under ROC curve under a nonignorable verification model, i.e., their method only requires either the disease model or the selection model to be correct. As a tradeoff to achieve the doubly robust property, the nonignorable parameter (the log odds ratio of verification for diseased versus non-diseased individuals) in their selection model is no longer identifiable. Hence they proposed to specify a value for the nonignorable parameter and conduct a sensitivity analysis. Fluss *et al.* (2009) applied the same methodology to construct the doubly robust sensitivity and specificity estimators, and hence the empirical ROC curve.

In practice, specifying the nonignorable parameter could be challenging because there is usually little prior information about the magnitude of the nonignorable parameter. So we may wish to estimate it from the data instead, under a parametric model. In this paper, we investigate the likelihood approach to estimate the nonignorable parameter. In order to carry out the estimation, we need to make some additional parametric assumptions for the disease regression model. We prove identifiability of the selection model with this additional parametric disease model. With the likelihood-based estimators of model parameters, we derive four forms of empirical estimators of the ROC curve as well as the Wilcoxon statistics for the area under ROC curve (AUC). We also prove the asymptotic normality for the AUC estimators. Simulation studies are conducted to compare our estimators of the AUC with doubly robust estimators.

Admittedly, the need for the nonignorability cannot be tested nonparametrically. We could only test the nonignorable parameter under some parametric selection model framework. Thus we recommend building up a plausible selection model based on the scientific knowledge about the sampling scheme, rather than rely on the statistical test.

Aside from selection models, the pattern mixture model (PMM) is also commonly used for analyzing nonignorable missing data (Little, 1993). However, we show that PMM is not identifiable in correcting nonignorable verification bias in Web Appendix A. More detailed discussions about the identifiability of PMM can be found in Little and Wang (1996). Our focus here is merely on the selection model.

This paper is organized as follows. In Section 2, we set up the selection model and disease model, as well as review the model settings in Rotnitzky, Faraggi, and Schisterman (2006). In Section 3, we prove model identifiability. We construct the estimators for the ROC curve and AUC in Section 4, and state the limit theorem for the estimated AUC. Simulation studies are conducted in Section 5, followed by a real example from the research in Alzheimer's disease in Section 6. Finally, we make concluding remarks in Section 7.

2. Model settings

Let V_i , D_i , T_i and X_i be the verification status, disease status, test results, and covariates of the i th patient, respectively, where $i = 1; 2; \dots, n$. We sometimes suppress the subscript i when there is no confusion. We assume the following selection model for the verification mechanism with a continuous test T :

$$\Pr(V_i=1|D_i, T_i, X_i) = \text{expit}(h(T_i, X_i; \beta) + \alpha D_i), \quad (1)$$

where $\text{expit}(x) \equiv \frac{\exp(x)}{1+\exp(x)}$ and α is the nonignorable parameter. For simplicity, we take $h(T_i, X_i; \beta) = \beta_0 + \beta_1 T_i + X_i^T \beta_2$, which is a natural consideration in practice. This can be easily extended to allow for interactions between T_i and X_i . If $\alpha = 0$, the missing mechanism is MAR; otherwise, it is nonignorable missing (NI). Note that the link function in (1) could also be Φ (standard normal cdf) for a probit model. We allow the link functions to be either expit or Φ in the next section, but we only use expit here for the purpose of easy interpretation. It follows from (1) that

$$\frac{\Pr(D=1|V=1, T, X)}{\Pr(D=1|V=0, T, X)} = \frac{\Pr(V=1|D=1, T, X)\Pr(D=1|T, X)/\Pr(V=1|T, X)}{\Pr(V=0|D=1, T, X)\Pr(D=1|T, X)/\Pr(V=0|T, X)} = \frac{\Pr(V=0|T, X)}{\Pr(V=1|T, X)} \exp(\beta_0 + \beta_1 T + X^T \beta_2 + \alpha). \quad (2)$$

Similarly,

$$\frac{\Pr(D=0|V=1, T, X)}{\Pr(D=0|V=0, T, X)} = \exp(\beta_0 + \beta_1 T + X^T \beta_2) \frac{\Pr(V=0|T, X)}{\Pr(V=1|T, X)}. \quad (3)$$

From (2) and (3), we have

$$\frac{\Pr(D=1|V=1, T, X)/\Pr(D=0|V=1, T, X)}{\Pr(D=1|V=0, T, X)/\Pr(D=0|V=0, T, X)} = \exp(\alpha). \quad (4)$$

The original interpretation of α based on the selection model is the log odds ratio of verification for a diseased subject versus a non-diseased subject. According to (4), α can also be interpreted as log odds ratio of being diseased for a verified case as compared to an unverified case with the same test result and covariates, as long as (1) holds. Then $\Pr(D=1|V=0, T, X)$ can be solved from (4) as a function of $\Pr(D=1|V=1, T, X)$,

$$\Pr(D=1|V=0, T, X) = \frac{\Pr(D=1|V=1, T, X)}{\exp(\alpha) + \Pr(D=1|V=1, T, X)(1 - \exp(\alpha))}. \quad (5)$$

As Rotnitzky, Faraggi, and Schisterman (2006) argued, the observed data are governed by $\Pr(D=1|V=1, T, X)$, $\Pr(V=1|T, X)$, $f(T, X)$. Hence if we changed α in (5) and kept $\Pr(D=1|V=1, T, X)$ the same, we would get a different disease probability for the unverified sample, while the observed (marginal) distribution still remains the same. Intuitively, since the disease model is restricted in the verified sample, $\Pr(D=1|V=0, T, X)$ could be any structure. Then they proposed to specify α and developed the doubly robust estimator for the area under ROC curve; that is, they showed that their AUC estimator is consistent if either $\Pr(V_i=1|D_i, T_i, X_i)$ or $\Pr(D=1|V=1, T, X)$ has a correctly specified parametric form.

Although the doubly robust property of their method is desirable, the validity of their method depends on the correct specification of the log odds ratio, α . It is quite a strong assumption to set α as known, and specifying α correctly could be challenging in practice.

In order to make α estimable, we need to assume a disease model for the whole sample,

$$\Pr(D_i=1|T_i, X_i)=\text{expit}(m(T_i, X_i;\gamma)). \quad (6)$$

For simplicity, we take $m(T_i, X_i;\gamma)=\gamma_0+\gamma_1 T_i+X_i^T \gamma_2$. The doubly robust property also has to be sacrificed as a cost to estimate α . Intuitively in this case, if we fix $\Pr(D=1|V=1, T, X)$ and change α in (5), $\Pr(D=1|V=0, T, X)$ will be changed and (6) may no longer hold in general. In other words, the restriction of (6) does provide some information in estimating α .

Let $\beta=(\beta_0, \beta_1, \beta_2^T)^T$ and $\gamma=(\gamma_0, \gamma_1, \gamma_2^T)^T$. Let $\pi_{di}=\Pr(V_i=1|D_i=d, T_i, X_i)$, given by the selection model (1), and $\rho_i=\Pr(D_i=1|T_i, X_i)$, given by the disease model (6). The log likelihood could be written as follows:

$$l(\alpha, \beta, \gamma)=\sum_i \{D_i V_i (\log \pi_{1i} \rho_i) + (1-D_i) V_i (\log \pi_{0i} (1-\rho_i)) + (1-V_i) \log (1-\pi_{1i} \rho_i - \pi_{0i} (1-\rho_i))\},$$

which has the same form as the multinomial likelihood. The three cells are given by $DV=1$, $(1-D)V=1$, and $V=0$, with the corresponding cell probabilities $\pi_{1i} \rho_i$, $\pi_{0i} (1-\rho_i)$ and $1-\pi_{1i} \rho_i - \pi_{0i} (1-\rho_i)$, respectively. If the expit link is used for the selection and disease model, then the score equations for α , β and γ are given as follows:

$$\begin{aligned} 0 &= \sum_i \left[D_i V_i (1-\pi_{1i}) - \frac{(1-V_i) \rho_i \pi_{1i} (1-\pi_{1i})}{1-\pi_{1i} \rho_i - \pi_{0i} (1-\rho_i)} \right], \\ 0 &= \sum_i \begin{pmatrix} 1 \\ T_i \\ X_i \end{pmatrix} \left[D_i V_i (1-\pi_{1i}) + (1-D_i) V_i (1-\pi_{0i}) - \frac{(1-V_i) (\rho_i \pi_{1i} (1-\pi_{1i}) + (1-\rho_i) \pi_{0i} (1-\pi_{0i}))}{1-\pi_{1i} \rho_i - \pi_{0i} (1-\rho_i)} \right], \\ 0 &= \sum_i \begin{pmatrix} 1 \\ T_i \\ X_i \end{pmatrix} \left[V_i (D_i - \rho_i) - \frac{(1-V_i) (\pi_{1i} - \pi_{0i}) \rho_i (1-\rho_i)}{1-\pi_{1i} \rho_i - \pi_{0i} (1-\rho_i)} \right]. \end{aligned}$$

The model parameters could be estimated by solving the above score equations using Newton-Raphson algorithm, or as suggested by previous literature, using an EM algorithm to facilitate the calculation.

As mentioned above, the likelihood function takes the form of the multinomial distribution, and the cell probabilities are determined by the disease and verification models. Although one could employ the Hosmer-Lemeshaw test to assess the goodness of fit, we do not recommend any formal test for the goodness of fit. The nonignorable verification model itself could not be tested nonparametrically from the data set, and hence the usefulness of the goodness of fit test may be in doubt. Instead, the disease and verification models should be based on scientific considerations, which may come from exogenous or prior knowledge about the disease and the sampling scheme.

3. Identifiability

With X and T given, the observed distribution is fully determined by two joint probabilities of D and V , $\Pr(D=1|V=1|T, X)$, which are functions of parameters α , β , and γ . Note that the two joint probabilities can be written as the product of $\pi_{d|j}$ and ρ_j , which are expit functions (or the normal cdf if using the probit model instead). Hence, $\Pr(D=1|V=1|T, X)$ and $\Pr(D=0, V=1|T, X)$ are simply products of two expit or Φ functions. We need to show that the two products are uniquely determined by the parameters for all possible T and X . For simplicity of notation, the vector of covariates, X , is omitted, since we can view X as fixed while varying T . Suppose T can take any real values.

We have the following lemma:

LEMMA 1: *Suppose $g(\cdot)$ is a strictly increasing and concave function with the support in $\mathbb{R}$. If*

$$g(a+bt)+g(c+dt)=g(a^*+b^*t)+g(c^*+d^*t) \quad (7)$$

holds for every $t \in \mathbb{R}$ and $b \neq d$, then $(a^, b^*, c^*, d^*) = (a, b, c, d)$ or (c, d, a, b) .*

Note that $a + bt$ can be seen as a straight line for t and that $a + bt$ could either parallel to or intersect with $a^* + b^*t$. Lemma 1 could be proved by considering all possible patterns of the straight lines $a + bt$, $c + dt$, $a^* + b^*t$, and $c^* + d^*t$. The detailed proof is given in Web Appendix A. Now we prove the parameter identification of the selection model (1) and disease model (6). Under our models, the logarithms of two probabilities, $\log \Pr(D=1, V=1|T, X)$ and $\log \Pr(D=0, V=1|T, X)$, are the sum of two $\log [\text{expit}(\cdot)]$ or $\log [\Phi(\cdot)]$ functions. Therefore Lemma 1 can be applied, and the theorem of identifiability is stated as follows.

THEOREM 1: *For a continuous scale test T , if $\Pr(D|T) \neq \Pr(D)$, i.e., the test result and the disease status are dependent, then the model specified in (1) and (6) is identifiable, i.e., the joint probabilities $\Pr(D=1, V=1|T=t)$ and $\Pr(D=0, V=1|T=t)$ are determined by a unique set of parameters.*

Note that Theorem 1 can easily be extended to allow a covariate, X . In addition, if T is ordinal but some covariate X is continuous, we could replace $T=t$ with $X=x$ in Theorem 1, and hence the selection model is still identifiable. In the next section, we see that the ROC curve and the area under ROC curve (AUC) are functions of the model parameters, α , β , and γ . Hence, the ROC curve and AUC would also be identifiable.

4. ROC curve and its area

When all study subjects had received disease verification, the true and false positive rates, $TPR(s)$ and $FPR(s)$, for a given cutoff point s , could be estimated from the empirical estimating functions:

$$S_1^* = \sum_i S_{1i}^* = \sum_i (I(T_i \geq s) - TPR(s))D_i, \quad (8)$$

$$S_2^* = \sum_i S_{2i}^* = \sum_i (I(T_i \geq s) - FPR(s))(1 - D_i), \quad (9)$$

where $I(\cdot)$ is the indicator function. We have

$$\widehat{TPR}(s) = \frac{\sum_i I(T_i \geq s) D_i}{\sum_i D_i}, \quad (10)$$

$$\widehat{FPR}(s) = \frac{\sum_i I(T_i \geq s) (1 - D_i)}{\sum_i (1 - D_i)}. \quad (11)$$

The empirical ROC curve is just a plot of $\widehat{TPR}(s)$ versus $\widehat{FPR}(s)$ for all possible s .

The area under ROC curve (AUC) is interpreted as the probability that a case has a larger test result than a control, plus half of the probability that a case and a control have the same test results, i.e.,

$$v = \Pr(T_2 > T_1 | D_2 = 1, D_1 = 0) + (1/2) \Pr(T_2 = T_1 | D_2 = 1, D_1 = 0).$$

Let $I_{ij} = I(T_i > T_j) + (1/2) I(T_i = T_j)$, where $I(\cdot)$ is the indicator function. In absence of missing data, the AUC could be estimated from the estimating functions based on U statistics:

$$U^* = \sum_{i \neq j} U_{ij}^* = \sum_{i \neq j} D_i (1 - D_j) (I_{ij} - v). \quad (12)$$

This leads to Wilcoxon statistic:

$$\widehat{v} = \left\{ \sum_{i \neq j} I_{ij} D_i (1 - D_j) \right\} / \left\{ \sum_{i \neq j} D_i (1 - D_j) \right\}.$$

With missing true disease status, we could estimate the empirical ROC curve and the AUC using the similar approach as in Alonzo and Pepe (2005). Denote $\hat{\pi}_i$ and $\hat{\rho}_i$ to be the estimated verification probability and disease probability, respectively. We propose four types of empirical estimators using $\hat{\pi}_i$ and $\hat{\rho}_i$.

The first approach is full imputation (FI), where D_i in the estimating functions (8), (9) and (12) are replaced by $\hat{\rho}_i$. In other words, all the disease statuses are imputed by the estimated ones, regardless whether they are observed or missing. The FI estimators for $\widehat{TPR}(s)$, $\widehat{FPR}(s)$, and the AUC are given as follows:

$$\begin{aligned} \widehat{TPR}_{FI}(s) &= \frac{\sum_i I(T_i \geq s) \hat{\rho}_i}{\sum_i \hat{\rho}_i}, \\ \widehat{FPR}_{FI}(s) &= \frac{\sum_i I(T_i \geq s) (1 - \hat{\rho}_i)}{\sum_i (1 - \hat{\rho}_i)}, \\ \widehat{v}_{FI} &= \left\{ \sum_{i \neq j} I_{ij} \hat{\rho}_i (1 - \hat{\rho}_j) \right\} / \left\{ \sum_{i \neq j} \hat{\rho}_i (1 - \hat{\rho}_j) \right\}. \end{aligned}$$

Denote $\rho_{0i} \equiv \Pr(D = 1 | V = 0, T, X)$ to be the disease probability given the disease status

being missing. Its estimated version 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 $\hat{\pi}_{1i} = \widehat{\Pr}(V_i = 1 | D_i = 1, T_i, X_i)$ and $\hat{\pi}_{0i} = \widehat{\Pr}(V_i = 1 | D_i = 0, T_i, X_i)$. The second approach is mean

score imputation (MSI), where only the missing D_i 's are replaced by $\hat{\rho}_{0i}$. Let $D_{MSI,i} = V_i D_i + (1 - V_i)\rho_{0i}$ and $\hat{D}_{MSI,i}$ be the estimated version with ρ_{0i} replaced by $\hat{\rho}_{0i}$. The estimated $TPR(s)$, $FPR(s)$, and the AUC are then given as follows:

$$\begin{aligned}\widehat{TPR}_{MSI}(s) &= \frac{\sum_i I(T_i \geq s) \hat{D}_{MSI,i}}{\sum_i \hat{D}_{MSI,i}}, \\ \widehat{FPR}_{MSI}(s) &= \frac{\sum_i I(T_i \geq s) (1 - \hat{D}_{MSI,i})}{\sum_i (1 - \hat{D}_{MSI,i})}, \\ \widehat{v}_{MSI} &= \left\{ \sum_{i \neq j} I_{ij} \hat{D}_{MSI,i} (1 - \hat{D}_{MSI,j}) \right\} / \left\{ \sum_{i \neq j} \hat{D}_{MSI,i} (1 - \hat{D}_{MSI,j}) \right\}.\end{aligned}$$

The third approach is inverse probability weighting (IPW). Each observation is weighted by the inverse of sampling fraction, i.e., the probability of getting disease verification. The corresponding estimators of $TPR(s)$, $FPR(s)$ and the AUC are:

$$\begin{aligned}\widehat{TPR}_{IPW}(s) &= \frac{\sum_i I(T_i \geq s) V_i D_i / \hat{\pi}_i}{\sum_i V_i D_i / \hat{\pi}_i}, \\ \widehat{FPR}_{IPW}(s) &= \frac{\sum_i I(T_i \geq s) V_i (1 - D_i) / \hat{\pi}_i}{\sum_i V_i (1 - D_i) / \hat{\pi}_i}, \\ \widehat{v}_{IPW} &= \left\{ \sum_{i \neq j} I_{ij} \frac{V_i V_j D_i (1 - D_j)}{\hat{\pi}_i \hat{\pi}_j} \right\} / \left\{ \sum_{i \neq j} I_{ij} \frac{V_i V_j D_i (1 - D_j)}{\hat{\pi}_i \hat{\pi}_j} \right\}.\end{aligned}$$

The fourth approach is pseudo doubly robust estimator (PDR). Let $D_{PDR,i} \equiv V_i D_i / \pi_i + (1 - V_i / \pi_i) \rho_{0i}$. The estimated $\hat{D}_{PDR,i}$ have π_i and ρ_{0i} replaced by $\hat{\pi}_i$ and ρ_{0i} . The estimated $TPR(s)$, $FPR(s)$ and AUC are then given as follows:

$$\begin{aligned}\widehat{TPR}_{PDR}(s) &= \frac{\sum_i I(T_i \geq s) \hat{D}_{PDR,i}}{\sum_i \hat{D}_{PDR,i}}, \\ \widehat{FPR}_{PDR}(s) &= \frac{\sum_i I(T_i \geq s) V_i (1 - \hat{D}_{PDR,i})}{\sum_i (1 - \hat{D}_{PDR,i})}, \\ \widehat{v}_{PDR} &= \left\{ \sum_{i \neq j} I_{ij} \hat{D}_{PDR,i} (1 - \hat{D}_{PDR,j}) \right\} / \left\{ \sum_{i \neq j} I_{ij} \hat{D}_{PDR,i} (1 - \hat{D}_{PDR,j}) \right\}.\end{aligned}$$

The functional form of the PDR estimator is the same as that in Rotnitzky, Faraggi, and Schisterman (2006). However, it does not have the doubly robust property in general, as the two probabilities $\hat{\pi}_i$ and $\hat{\rho}_i$ are estimated together from the likelihood function, and correct specification of both the verification model and the disease model is required.

The empirical ROC curve using FI and MSI estimators is monotonic, because $\hat{\rho}_i$ and $\hat{D}_{MSI,i}$ are both between 0 and 1. The IPW estimator of the ROC curve is also monotonic, because $V_i D_i / \hat{\pi}_i$ and $V_i (1 - D_i) / \hat{\pi}_i$ are both positive. However, the PDR estimator of the ROC curve may be nonmonotonic, as $\hat{D}_{PDR,i}$ could be outside the $[0, 1]$ interval. One can use isotonic regression techniques for a correction as suggested by Fluss *et al.* (2009), but we do not elaborate the details here.

We note that $D_{MSI,i}$ and $D_{PDR,i}$ are both unbiased estimates of ρ_i . We have

$$\begin{aligned}
& E(D_{MSI,i} | T_i, X_i) \\
&= E(V_i D_i \\
&\quad + (1 - V_i) \rho_{0i} | T_i, X_i) \\
&= \Pr(V_i \\
&= 1 | T_i, X_i) E_{D_i | V_i=1, T_i, X_i}(V_i D_i \\
&\quad + (1 - V_i) \rho_{0i}) \\
&+ \Pr(V_i = 0 | T_i, X_i) E_{D_i | V_i=0, T_i, X_i}(V_i D_i \\
&\quad + (1 - V_i) \rho_{0i}) \\
&= \Pr(V_i \\
&= 1 | T_i, X_i) \Pr(D_i \\
&= 1 | V_i \\
&= 1, T_i, X_i) \\
&+ \Pr(V_i \\
&= 0 | T_i, X_i) \rho_{0i} \\
&= \Pr(D_i \\
&= 1 | T_i, X_i) = \rho_i
\end{aligned}$$

and

$$E(D_{PDR,i} | T_i, X_i) = E(V_i D_i / \pi_i + (1 - V_i / \pi_i) \rho_{0i} | T_i, X_i) = E_{D_i | T_i, X_i} \left[E_{V_i | D_i, T_i, X_i} (V_i D_i / \pi_i + (1 - V_i / \pi_i) \rho_{0i}) \right] = E_{D_i | T_i, X_i} D_i = \rho_i.$$

THEOREM 2: *The FI, MSI, IPW, and PDR estimators of $TPR(s)$, $FPR(s)$ and AUC are all consistent.*

The proof of Theorem 2 is sketched in Web Appendix A. In fact, as long as a predicted $\hat{D}_i$ has expectation ρ_i , we could impute the gold standard with $\hat{D}_i$ and attain a consistent estimator of the ROC curve and the AUC.

The asymptotic normality for the estimated $TPR(s)$ and $FPR(s)$ can be shown similarly as in the Appendix A of Alonzo and Pepe (2005). The proof was based on estimating equation theories as well as multivariate delta method. As the confidence ellipse of $TPR(s)$ and $FPR(s)$ is not our main interest in this paper, we do not elaborate the results here.

Denote the model parameters $\theta^T = (\alpha, \beta^T, \gamma^T)$. The estimating function of the AUC for FI, MSI, IPW and PDR estimators are denoted by

$$\begin{aligned}
U_{ij}^{FI}(\nu, \theta) &\equiv \rho_i(1 - \rho_j)(I_{ij} - \nu), \\
U_{ij}^{MSI}(\nu, \theta) &\equiv D_{MSI,i}(1 - D_{MSI,j})(I_{ij} - \nu), \\
U_{ij}^{IPW}(\nu, \theta) &\equiv \frac{V_i V_j D_i (1 - D_j)}{\pi_i \pi_j} (I_{ij} - \nu), \\
U_{ij}^{PDR}(\nu, \theta) &\equiv D_{PDR,i}(1 - D_{PDR,j})(I_{ij} - \nu).
\end{aligned}$$

For simplicity, we denote these estimating functions with $U_{ij}(\nu, \theta)$. Let

$$Q_i(v, \theta) \equiv E_i[U_{ij}(v, \theta) + U_{ji}(v, \theta)] + \left[E \frac{\partial}{\partial \theta} U_{ji}(v, \theta) \right] I(\theta)^{-1} \dot{l}_i(\theta), \quad (13)$$

where E_j is the expectation with respect to (V_j, D_j, T_j, X_j) , $l_i(\theta)$ is the i th subject's contribution to the score function, and $I(\theta) \equiv -E \frac{\partial}{\partial \theta} \dot{l}_i(\theta)$ is the information matrix for θ . Let

$$\hat{Q}_i \equiv n^{-1} \left[\sum_{j=1}^n U_{ij}(\hat{v}, \hat{\theta}) + U_{ji}(\hat{v}, \hat{\theta}) \right] - n^{-1} \left[\sum_{i=1}^n \sum_{j=1, j \neq i}^n \frac{\partial}{\partial \theta} U_{ij}(v, \theta) \Big|_{v=\hat{v}, \theta=\hat{\theta}} \right] \times \left[\sum_{i=1}^n \frac{\partial}{\partial \theta} \dot{l}_i(\theta) \Big|_{\theta=\hat{\theta}} \right]^{-1} \dot{l}_i(\hat{\theta}),$$

where $\hat{v}$ is either $\hat{v}_{FB}$, $\hat{v}_{MSF}$, $\hat{v}_{IPW}$ or $\hat{v}_{PDR}$. The following theorem states the asymptotic normality for $\hat{v}$ under some regularity conditions given in Web Appendix A.

THEOREM 3: *Suppose the verification model (1), the disease model (6), and the regularity conditions r1-r8 in Web Appendix A hold, then the asymptotic result is stated as follows:*

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

, where

$$\Omega = \frac{\text{var}[Q_i(v, \theta)]}{[\Pr(D_i=0)\Pr(D_i=1)]^2}.$$

In practice $\text{var}[Q_i(v, \theta)]$ can be consistently estimated by $n^{-1} \sum_{i=1}^n \hat{Q}_i^2$. The disease prevalence are estimated by the denominator of the corresponding $TPR(s)$ and $FPR(s)$ estimators. A sketch of proof of the above theorem can be found in Web Appendix A. Intuitively, the first part of $Q_i(v, \theta)$ in (13), $E_j[U_{ij}(v, \theta) + U_{ji}(v, \theta)]$, represents the variation

due to empirical estimation of Wilcoxon statistic; the second part, $\left[E \frac{\partial}{\partial \theta} U_{ji}(v, \theta) \right] I(\theta)^{-1} \dot{l}_i(\theta)$, represents the variation due to estimation of θ by $\hat{\theta}$

Recently, He, Lyness and McDermott (2009) derived the asymptotic variance for IPW estimator of AUC under ignorable verification bias. Their result is a special case of the

above theorem. When the missingness is ignorable, the IPW estimator has $E \frac{\partial}{\partial \theta} U_{ji}(v, \theta) = 0$ hence the variability due to estimating θ disappears. The asymptotic variance is simply the variability from using the U-statistic, as shown in He, Lyness and McDermott (2009).

In our model settings, the parameters in the selection model and in the disease model have to be estimated together from the score equations. Our procedure requires both the selection model and disease model to be true. On the other hand, in Rotnitzky, Faraggi, and Schisterman (2006), the parameters for their selection model and disease model could be estimated separately due to the known nonignorable parameter α . The tradeoff to estimate α is to give up the doubly robust property.

5. Simulations

In this section, we report the primary simulation results on comparison of the proposed estimates of AUC (FI, MSI, IPW and PDR estimates) with the doubly robust estimates (DR) in Rotnitzky, Faraggi, and Schisterman (2006). Some additional simulations are shown in Web Appendix B. We set the sample size as $n = 5000$. For each subject, we generated a continuous covariate from the standard normal distribution, $X_1 \sim N(0, 1)$, a binary covariate from a bernoulli distribution, $X_2 \sim \text{Bernoulli}(0.5)$, and the test result from a uniform distribution, $T \sim U(-1, 1)$. Four different scenarios of selection and disease models were considered:

Scenario A: The disease model (6) with $m(T, X; \gamma) = 1.5X_1 + X_2 + 2.5T$ and the selection model (1), with $h(T, X; \beta) = -2 - 0.8X_1 - 0.7X_2 + T$ and $\alpha = 1.2$. The correct models were fitted.

Scenario B: The disease model (6) with $m(T, X; \gamma) = 1.5X_1 + X_2 + 2.5T$ and the selection model (1), with $h(T, X; \beta) = -1.3 - 0.8X_1 - 0.7X_2 + T$ and $\alpha = 0$. The missingness was ignorable but we still estimated α as in the situation of nonignorable missingness.

Scenario C: The same disease model and selection model as in Scenario A. The fitted disease and selection models use $X_1^{1/3}$ instead of X_1 . (In this setting, the true covariate's effect is cubic while a linear covariate's effect is fitted.)

Scenario D: The disease model (6) with $m(T, X; \gamma) = 1.5X_1 + X_2 + 2.5T + 0.4X_1X_2$ and the selection model (1), with $h(T, X; \beta) = -2 - 0.8X_1 - 0.7X_2 + T + 0.5X_1X_2$ and $\alpha = 1.2$. The true models had interaction terms of the covariates but we only fitted the main effects in the estimation.

In all the four scenarios the disease prevalence was about 57% and the verification proportion was about 20%. The true AUC was 0.7563 for the first three scenarios and 0.7458 for the last scenario. DR estimates with $\alpha = -0.6, 0, 0.6, 1.2, 1.8$ were considered as a comparison. Table 1 shows the results from 500 simulations.

Comparing the four proposed empirical estimates of the AUC, IPW and PDR estimates are generally less efficient than FI and MSI estimates. In scenario A, all the four proposed AUC estimates are consistent, while the DR estimates with misspecified α could lead to serious bias. In this setting, FI and MSI estimates are more efficient than DR estimates with correct specification of $\alpha = 1.2$. This is probably because the DR estimate contains the term of $V/\hat{\pi}_j$ and its variability is inflated when $\hat{\pi}_j$ is small. The same reason may account for the large standard errors of the IPW and PDR estimates. The coverage rates of the four proposed AUC estimates are close to the 95% nominal level. In scenario B, when the missingness is truly ignorable, our proposed estimates all have small bias, but FI and PDR estimates seem to have slightly lower CI coverage rate. The DR estimate with correct specification of $\alpha = 0$ performs well, but the DR estimates with misspecified α still lead to biased results. Scenario C and D both examines the influence of misspecification of the disease and verification models. Scenario C may represent a minor departure from the true model, and the coverage rates for our proposed estimates are close to 90%. With a greater departure from the true model in Scenario D, our proposed estimates still give reasonable results with coverage rates above 80%, showing some robust property of the proposed estimates. It would be interesting for future investigation why the FI and MSI estimates generally perform better than IPW and PDR estimates under model misspecification.

The above simulation results show that our proposed AUC estimates are comparable to the DR estimate with the correctly specified α . FI and MSI estimates are somewhat more efficient than the DR estimate, while IPW and PDR estimates are less efficient. Although DR estimate is robust against misspecification of either disease model or verification model, it is usually highly sensitive to the nonignorable parameter α . Among FI, MSI, IPW and PDR estimates, we would recommend FI and MSI estimate in practice because of its efficiency.

6. Example: NACC data

We illustrated our proposed method with the data collected by National Alzheimer's Coordinating Center (NACC), which consist of 74,075 patients. The Mini Mental State Examination (MMSE) was given to about 80% of the patients in the database. The test can be scored from 0 to 30, with lower scores indicating more severe cognitive impairment. Our goal is to evaluate the diagnostic accuracy of the MMSE test in detecting Alzheimer's disease (AD). The gold standard for AD is based on Reagan Criteria (Bennett *et al.*, 2006), and requires brain autopsy. So people still alive cannot have disease verification, which is the primary reason of missing disease status. Another reason may be that the patients or their family opt not to have brain autopsy. As a result, the disease verification is only available to about 11% of the patients. Due to the fact that living people may have better health status and hence are less likely to have AD, the missingness of the disease status is probably nonignorable. Other available covariates, which are believed to be related to either the probability of being demented or the disease verification, include age at the MMSE test, gender, race, marital status, clinical diagnosis of AD, other disease conditions (i.e., stroke, Parkinson's disease, depression). Patients with missing test results or missing covariates were excluded from our analysis, which resulted in 53,063 patients in total. We denote T to be the MMSE test, D to be the true AD status, and X to be the above patient covariates. Table 2 shows the summary statistics for all the patients used in our analysis. Figure 1 displays the distribution of the MMSE test for all the patients and for stratified verification and disease status. We can see that subjects with lower MMSE scores are more likely to get the disease verification.

We used the selection model (1) with

$$\Pr(V=1|D, T, X) = \text{expit}(\beta_0 + X^T \beta_1 + \beta_2 T + \alpha D),$$

and the disease model (6) with

$$\Pr(D=1|T, X) = \text{expit}(\gamma_0 + X^T \gamma_1 + \gamma_2 T),$$

where X was the vector of all the aforementioned covariates. In the above models, we employed a linear transformation on age and test result by considering $(age - 70)/10$ and $(30 - T)/15$, just for the convenience of reporting their coefficients. The estimated parameters and their standard errors are given in Table 3. We find that all variables except clinical AD status are significant in the selection model. The nonignorable parameter is estimated to be 0.718, indicating that the odds ratio of verification for the diseased versus the non-diseased is 2.05 (95% CI: 1.45, 2.91). The disease model identifies some significant risk factors such as age, gender, clinical AD diagnosis, as well as the test result. In fact, our main interest here is not on the risk factors, but on the predicted individual verification probability and disease probability. Hence we just keep the predetermined model that is scientifically plausible, rather than considering any model selection.

We also try different predictors in the models, but it seems that the nonignorable parameter does not change much. For example, when deleting the race variable in the disease model, α is estimated to be 0.721 (95% CI: 0.375, 1.067); when adding the interaction between T and gender, $\hat{\alpha} = 0.696$ (95% CI: 0.331, 1.060); when adding the interaction between age and gender, $\hat{\alpha} = 0.759$ (95% CI: 0.418, 1.100). Here we just report the models with main effects of the covariates, and calculate the AUC thereafter.

Figure 2 shows the estimated ROC curves using FI, MSI, IPW and PDR methods. We can see that the FI and MSI estimates of the ROC curve virtually overlap with each other. The IPW curve is a little lower than the FI curve, and the PDR curve is slightly higher than the FI curve. As for the AUC, the FI estimate for AUC is 0.7350 (95% CI: 0.7215, 0.7484); the MSI estimate for AUC is 0.7355 (95% CI: 0.7243, 0.7468); the IPW estimate for AUC is 0.7164 (95% CI: 0.6983, 0.7344); the PDR estimate for AUC is 0.7514 (95% CI: 0.7340, 0.7688). Similarly as in the simulation studies, FI and MSI estimates have narrower confidence intervals than the IPW and PDR estimates. The reason why IPW estimates are not as precise as the FI and MSI estimates may be due to the low verification probability. With only about 10% of verification, the inverse of selection probability would be large for many subjects, so as to make the IPW estimates less stable and precise. In addition, the DR estimates with different $\alpha \in [-2, 2]$ are plotted in Figure 3, as well as the FI, MSI and IPW estimates. The 95% CI's for DR estimates are much wider than the FI and MSI estimates. This is also due to the small verification probabilities, as DR estimates make use of the inverse of the estimated verification probability.

7. Discussion

In this paper, we have proposed a likelihood-based approach to estimate the ROC curve and AUC under nonignorable verification bias. We have proved that the selection model (1) together with the disease model (6) is identifiable and hence can be estimated by solving the score equations. We have also proposed the FI, MSI, IPW and PDR estimators for the empirical ROC curve and AUC. Compared to the doubly robust estimators proposed by Rotnitzky, Faraggi, and Schisterman (2006), we need to specify the parametric disease model (6) and give up the doubly robust property in order to estimate the nonignorable parameter α . With the parametric setting of the joint probability of D and V , the selection model becomes identifiable. The identifiability basically comes from the disease model on the whole population.

For real-world applications, it could be challenging to make both the disease model and the verification model exactly right. However, if we could collect necessary information on the reasons of missing gold standard and disease mechanism, it is still possible to build approximately correct models from the scientific point of view. In practice, logistic regressions often provide reasonably good approximations, and are easy to interpret. Like in linear regressions, one has to “extrapolate” the associations between the variables in a way that is both parsimonious and scientifically plausible. We use the logistic regression models in both simulations and the real example, which is easy to implement and interpret. The covariates are all selected based on scientific knowledge in the literature, and we report the models with main effects only.

Admittedly, a reasonable large sample size (usually several thousand) is required to avoid model failure. These kinds of examples do exist in epidemiological studies: a large number of subjects can be screened for a disease by a simple and inexpensive test; but the gold standard could only be obtained for a small proportion due to its complexity or high cost. The NACC data serves as a good example of such studies with only 11% of the patients having disease verification. Despite of much missing information, we are still able to

estimate the nonignorable selection model. In other situations where only a very small sample of patients could be recruited anyway, one has to make stronger MAR assumptions.

Molenburghs and Kenward (2007) pointed out that “every NMAR model has a MAR bodyguard” in their section 19.2. They factor out the likelihood in terms of the pattern mixture model (PMM) and show that some MAR model could achieve the same fit to the data as the NMAR model. In fact, this is another way to say that the nonignorable missingness assumption is untestable from the data alone. In a sense, specifying a MAR model under PMM has the same difficulty as specifying the NMAR model. MAR model has to extrapolate that $\Pr(D|V=0, T, X) = \Pr(D|V=1, T, X)$, which is again untestable, as one has no information about the missing disease prevalence from the data alone. Therefore, we may resort to some prior knowledge about the data collection procedures, and infer about the missing data mechanism. We suggest that one should build both selection and disease models from the scientific point of view, and the feasible model(s) should be determined prior to data analysis. Any data-driven model selection procedures are no longer valid under NMAR model.

We also note that, without any assumptions on $\Pr(T_i = d|X_i)$, we can only obtain the marginal ROC curve, not the covariate-specific one. This is because the covariate-specific ROC curve

depends on the probability $\Pr(T|D, X) = \frac{\Pr(D|T, X)\Pr(T|X)}{\int \Pr(T|X)\Pr(D|T, X)dT}$, which cannot be identified without $\Pr(T|X)$.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

The authors want to thank Dr. Thomas Koepsell and Dr. Walter Kukull for the helpful discussions and their insightful comments. The work was supported in part by NIH/NIA grant U01AG016976 and National Science Foundation of China grant NSFC30728019. 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. *Applied Statistics*. 2005; 54:173–190.
- Baker SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics*. 1995; 51:330–337. [PubMed: 7539300]
- Begg CB. Biases in assessment of diagnostic tests. *Statistics in Medicine*. 1987; 6:411–423. [PubMed: 3114858]
- 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, Arvantitakis 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]
- 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]

- Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*. 2003; 59:163–171. [PubMed: 12762453]
- Little RJA. Pattern-mixture models for multivariate incomplete data. *Journal of the American Statistical Association*. 1993; 88:125–134.
- Little RJA, Wang Y. Pattern-mixture models for multivariate incomplete data with covariates. *Biometrics*. 1996; 52:98–111. [PubMed: 8934587]
- Molenburghs, G.; Kenward, M. *Missing data in Clinical Studies*. New York: John Wiley; 2007.
- Rodenburg CA, Zhou XH. ROC curve estimation when covariates affect the verification process. *Biometrics*. 2000; 56:1256–1262. [PubMed: 11129488]
- 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.
- van der Vaart, A. *Asymptotic statistics*. Cambridge University Press; 1998.
- Zhou XH. A nonparametric ML estimate of an ROC curve area corrected for verification bias. *Biometrics*. 1996; 52:310–316.
- Zhou XH. Comparing correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias. *Biometrics*. 1998; 54:349–366.
- 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.
- Zhou XH, Castelluccio P. Adjusting for nonignorable verification bias in clinical studies for Alzheimer's disease. *Statistics in Medicine*. 2004; 23:221–230. [PubMed: 14716724]
- Zhou XH, Rodenburg CA. Estimating an ROC curve in the presence of nonignorable verification bias. *Communications in Statistics*. 1998; 27:635–657.

Figure 1.
MMSE distribution by AD status and verification.

Figure 2.
Estimated ROC curve using FI, MSI, IPW and PDR approaches.

Figure 3. DR estimates of the AUC with different α values, as compared to FI, MSI, IPW and PDR estimates with associated 95% CI's.

Table 1

Comparison of the proposed FI, MSI, IPW and PDR estimates with DR estimates.

	Estimate	Bias (%) ^a	SD ^b	SE ^c	Coverage (%)
Scenario A	FI	-0.0	16.0	16.5	94.6
	MSI	-0.0	16.0	17.5	95.2
	IPW	-0.1	24.3	25.1	95.8
	PDR	0.0	23.6	22.1	93.0
	DR ($\alpha=1.2$) ^d	-0.2	21.7	20.8	93.6
	DR ($\alpha=-0.6$)	-4.4	19.3	19.6	61.6
	DR ($\alpha=0$)	-2.7	19.5	19.6	82.0
	DR ($\alpha=0.6$)	-1.3	20.3	20.0	90.0
	DR ($\alpha=1.8$)	0.8	23.7	21.8	92.0
Scenario B	FI	-0.0	19.2	18.8	91.2
	MSI	-0.0	19.2	22.8	93.6
	IPW	-0.2	27.2	27.3	94.4
	PDR	-0.0	24.0	21.1	91.8
	DR ($\alpha=0$) ^d	0.0	19.4	19.3	94.8
	DR ($\alpha=-0.6$)	-1.6	19.3	19.2	90.2
	DR ($\alpha=0.6$)	1.3	20.1	19.8	91.4
	DR ($\alpha=1.2$)	2.3	21.4	20.5	84.2
	DR ($\alpha=1.8$)	3.0	23.3	21.4	79.2
Scenario C	FI	0.3	18.8	16.9	90.8
	MSI	0.3	18.8	21.2	89.8
	IPW	1.6	27.3	26.9	90.8
	PDR	2.3	29.3	29.0	93.8
	DR ($\alpha=1.2$) ^d	0.0	22.2	22.0	93.0
	DR ($\alpha=-0.6$)	-4.1	19.5	19.6	66.2
	DR ($\alpha=0$)	-2.5	19.7	19.9	84.0
	DR ($\alpha=0.6$)	-1.2	20.8	20.8	90.2
	DR ($\alpha=1.8$)	1.2	23.5	23.0	90.0

Scenario D	Estimate	Bias (%) ^a	SD ^b	SE ^c	Coverage (%)
	FI	-2.0	21.8	19.7	91.0
	MSI	-2.0	21.8	24.3	94.8
	IPW	-3.6	27.8	26.6	83.8
	PDR	-2.4	24.3	20.4	81.0
	DR ($\alpha = 1.2$) ^d	-1.6	20.4	19.2	88.0
	DR ($\alpha = -0.6$)	-4.9	19.3	19.0	52.4
	DR ($\alpha = 0$)	-3.5	19.0	18.5	70.0
	DR ($\alpha = 0.6$)	-2.5	19.3	18.6	81.2
	DR ($\alpha = 1.8$)	-0.8	22.1	20.1	90.8

^aBias as a percentage of the true AUC.
^b1000 times the empirical standard deviation of the point estimates.
^c1000 times the average of the estimated standard error.
^dDR estimate with correct specification of α .

Table 2

Summary statistics of the NACC data.

Variable	Summary statistics
<i>D</i> (gold standard)	AD: 3.00%; AD-free: 7.29%; missing: 89.71%
<i>T</i> (MMSE)	mean: 21.5; SD: 7.7; median: 24
age	mean: 72.6; SD: 10.6; median: 74.2
gender	male: 39.8%; female: 60.2%
race	white: 81.9%; other: 18.1%
marital status	married: 56.8%; other: 43.2%
clinical diagnosis of AD	yes: 53.0%; no: 47.0%
stroke	yes: 11.9%; no: 88.1%
Parkinson's disease	yes: 4.4%; no: 95.6%
depression	yes: 16.4%; no: 83.6%

Table 3

Estimated model parameters and their standard errors. (Significant coefficients are in bold font.)

	Verification model	Disease model
Intercept	-4.527 (0.089)	-1.101 (0.252)
age	0.086 (0.017)	0.134 (0.034)
gender	0.587 (0.037)	-0.468 (0.075)
race	1.696 (0.070)	-0.025 (0.175)
marital status	-0.195 (0.035)	0.129 (0.080)
clinical AD diagnosis	0.079 (0.083)	1.881 (0.070)
stroke	0.305 (0.043)	-0.100 (0.095)
Parkinson's disease	0.641 (0.058)	0.234 (0.122)
depression	-0.202 (0.044)	0.110 (0.099)
<i>T</i>	0.203 (0.040)	0.784 (0.071)
<i>D</i>	0.718 (0.178)	-