REVSTAT – Statistical Journal
Volume 12, Number 1, March 2014, 67–83

VERIFICATION BIAS—IMPACT AND METHODS FOR CORRECTION WHEN ASSESSING ACCURACY OF DIAGNOSTIC TESTS

Author: Todd A. Alonzo
– Department of Biostatistics, University of Southern California, Los Angeles, CA, USA
talonzo@childrensoncologygroup.org

Abstract:

- Sometimes it is not feasible to obtain disease status verification for all study subjects. Analysis of only those with disease ascertainment can result in biased estimates of the accuracy (sensitivity, specificity, ROC curve) of a diagnostic test, screening test, or biomarker if the estimation method does not properly account for the missing disease ascertainment. This paper discusses the impact of this bias, verification bias, when estimating the accuracy of dichotomous and continuous diagnostic tests. In addition, methods to correct for verification bias are described. Areas that require additional attention are also highlighted.

Key-Words:

- *imputation; inverse probability weighting; ROC curve; sensitivity; specificity.*

AMS Subject Classification:

- 62F10, 62F15, 62J12, 62P10.

1. INTRODUCTION

Estimating accuracy of a diagnostic test, screening test, or biomarker is ideally done by determining disease status using a gold standard test or reference test for all study subjects. However, sometimes disease status verification via the reference test is not obtained for all study subjects because the reference test is too costly or invasive to be applied to all study subjects. When this is the case, subjects who appear to be at high risk may be more likely to have disease status assessed via the reference test than those who appear to be at lower risk. Analysis of only those with disease ascertainment can result in biased estimates of accuracy if the estimation methods do not properly account for nonrandom disease ascertainment. This bias is known as work-up bias (Ransohoff and Feinstein, 1978) and verification bias (Begg and Greenes, 1983). Verification bias can yield investigators to incorrectly conclude that a diagnostic test is more accurate than it is or the reverse that the test is less accurate than it actually is. This can have significant implications if the diagnostic test is implemented in practice based on incorrect conclusions.

Incomplete disease verification can occur by design or be unplanned. As expected, designed partial verification is more likely to occur in prospective studies while retrospective studies more typically have unplanned partial verification. In some studies it is not feasible to obtain the reference standard on subjects thought to be at low risk so the study is designed with partial verification. For example, the Prostate Cancer Prevention Study (Thompson *et al.*, 2005) of the effects of prostate specific antigen (PSA) the reference standard, prostate biopsy, was recommended only if the PSA level was greater than 4.0 ng/ml or rectal examination result was abnormal.

Methods for assessing accuracy of diagnostic tests differ depending on how the test is measured. Diagnostic tests can yield dichotomous results indicating presence or absence of particular condition or disease. For example, stress echocardiography to detect significant coronary artery stenosis. Diagnostic tests can also yield results that are measured on a continuous scale, such as, prostate specific antigen (PSA) for detecting prostate cancer. Typically, sensitivity, specificity, positive predictive value (PPV), and negative predictive value (NPV) are used to assess the accuracy of dichotomous diagnostic tests. Conversely, receiver operating characteristic (ROC) curves and corresponding summary measures, such as area under the ROC curve (AUC), are used to assess the accuracy of continuous tests.

Correcting for verification bias can be framed as a missing data problem where true disease status is missing for a subset of study subjects. Each approach for bias correction makes an assumption about the mechanism for the missing-

ness of disease verification (Little and Rubin, 1987). Disease status is considered missing completely at random (MCAR) if disease verification is independent of observed and unobserved data. Disease status is considered missing at random (MAR) when disease verification is only a function of observed data and is considered nonignorable (NI) when disease verification depends on unobserved data.

In Section 2 the notation for this paper is introduced. Sections 3 and 4 discuss the impact of verification bias when estimating the accuracy of a dichotomous diagnostic test and a continuous diagnostic test, respectively, and summarize available bias correction methods. We end with a Discussion.

2. NOTATION

Consider a study with n subjects on which the diagnostic test T is measured. Let D be disease status, as measured by a gold standard or reference test, where $D = 1$ corresponds to presence of disease and $D = 0$ corresponds to absence of disease. Further, let V be verification status where $V = 1$ if disease status is verified and $V = 0$ otherwise. There are n_V subjects with disease verification and $n_{\bar{V}} = n - n_V$ without disease verification.

3. DICHOTOMOUS TEST

Consider a dichotomous test T where $T = 1$ indicates a positive test and $T = 0$ indicates a negative test. Table 1 summarizes the observed data from a study of $n = n_1 + n_0$ subjects in which disease verification is not obtained in u_1 test positives and u_0 test negatives.

Table 1: Observed data for the verification bias problem when T is dichotomous.

V	D	$T = 1$	$T = 0$
1	1	s_1	s_0
1	0	r_1	r_0
0	Missing	u_1	u_0
Total:		n_1	n_0

3.1. Impact of bias

Consider a study of 1000 subjects to assess the sensitivity and specificity of a dichotomous screening test with a true sensitivity of 80%, true specificity of 90%, and disease prevalence, $P(D = 1)$, of 10%. Data from this hypothetical study are summarized on the left-hand side of Table 2. If the study design is such that disease verification is obtained for all subjects who test positive and only 10% of subjects who test negative, this can result in observing the data on the right-hand side of Table 2.

Table 2: Left side: results when disease verification is obtained for everyone. Right side: observed data when disease verification is obtained for all subjects who test positive and only 10% of subjects who test negative.

V	D	$T = 1$	$T = 0$
1	1	80	20
1	0	90	810
0	Missing	0	0
Total:		170	830

V	D	$T = 1$	$T = 0$
1	1	80	2
1	0	90	81
0	Missing	0	747
Total:		170	830

If we only consider test results for those with disease verification, referred to as complete case estimators, the observed sensitivity is $s_1/(s_1 + s_0) = 80/82$ or 98% and the observed specificity is $r_0/(r_0 + r_1) = 81/171$ or 47%. This illustrates that if test positives are more likely to receive disease verification than test negatives, observed sensitivity overestimates true sensitivity (98% *vs.* 80%) and observed specificity underestimates true specificity (47% *vs.* 90%). This verification bias can cause investigators to make incorrect conclusions regarding the accuracy of a test under evaluation.

It can be shown that PPV, $P(D = 1 \mid T = 1)$, is 47% using the full data and 47% using only those who received disease verification. Similarly, NPV, $P(D = 0 \mid T = 0)$ is 98% using the full data and also when only those who received disease verification are used. There is no bias in the complete case estimators of PPV and NPV because disease verification is only a function of the test results T, and PPV and NPV are, by definition, calculated conditional on T. See Zhou (1994) for a detailed discussion of the effect of verification bias on positive and negative predictive values. Next, we discuss methods to correct for the biased sampling when estimating sensitivity and specificity.

3.2. Bias correction methods

3.2.1. MAR approaches

Begg and Greenes (1983) developed a bias correction method for sensitivity and specificity by using Bayes' Rule and assuming disease status is MAR. First, consider estimating the sensitivity of a test. Bayes' Rule can be used to re-write sensitivity as

$$\begin{aligned} P(T=1 \mid D=1) &= \frac{P(T=1, D=1)}{P(D=1)} \\ &= \frac{P(D=1 \mid T=1)\, P(T=1)}{P(D=1 \mid T=1)\, P(T=1) + P(D=1 \mid T=0)\, P(T=0)} \,. \end{aligned} \tag{3.1}$$

Each quantity on the right-hand-side of (3.1) can be directly estimated from the observed data using empirical estimates. In particular, $P(T)$ can be estimated using data from all subjects, and $P(D \mid T)$ can be estimated using the verification group since by the MAR assumption $P(D \mid T) = P(D \mid T, V = 1)$. Substituting empirical estimates of the probabilities in (3.1) results in the following unbiased estimate of sensitivity

$$\widehat{P}(T=1 \mid D=1) = \frac{\frac{s_1 n_1}{s_1 + r_1}}{\frac{s_1 n_1}{s_1 + r_1} + \frac{s_0 n_0}{s_0 + r_0}} \,. \tag{3.2}$$

A bias-corrected estimate of specificity can be calculated in a similar fashion.

$$\widehat{P}(T=0 \mid D=0) = \frac{\frac{r_0 n_0}{s_0 + r_0}}{\frac{r_0 n_0}{s_0 + r_0} + \frac{r_1 n_1}{s_1 + r_1}} \,. \tag{3.3}$$

It can be shown that these estimators of sensitivity and specificity are maximum likelihood estimators. Furthermore, this approach can be considered single imputation as compared with multiple imputation which is discussed later. The delta method can be used to develop variance estimators for sensitivity and specificity.

Iglesias-Garriz *et al.* (2005) performed a study to estimate the sensitivity and specificity of stress echocardiography to detect significant coronary artery disease (CAD). The study involved 487 consecutive patients presenting at a hospital emergency room with nontraumatic chest pain, and who were administered stress echocardiography. Table 3 presents a tabulation of the study data, where using our notation T represents stress echocardiography, D is CAD, and V is an indicator of whether CAD status was determined. Of the 487 patients with stress echocardiography results, only 78 (16%) received disease verification via

coronary angiography to determine presence or absence of CAD. Furthermore, a higher percentage of those who tested positive with stress echocardiography received disease verification than those who tested negative with stress echocardiography (62.5% *vs.* 6.9%).

Table 3: Tabulation of Stress Echocardiography (T), CAD status (D), and disease verification status (V) in the study by Iglesias-Garriz *et al.*

V	D	$T=1$	$T=0$
1	1	43	15
1	0	7	13
0	Missing	30	379
Total:		80	407

Using only those with CAD status obtained, the complete case estimate of sensitivity is 74.1% (43/58) and the complete case estimate of specificity is 65.0% (13/20). Applying Equations 3.2 and 3.3, the Begg and Greenes estimate of sensitivity is 24.0% and corrected estimate of specificity is 94.4%. In this study, the uncorrected estimate of sensitivity clearly overestimates the corrected estimate while the uncorrected specificity substantially underestimates the corrected estimate.

Harel and Zhou (2006) discuss the use of multiple imputation to estimate sensitivity and specificity of a binary diagnostic test in the presence of verification bias. Each missing disease status is replaced by M imputed values and then each of the M complete data sets is analyzed using complete data methods. The M point estimates of sensitivity and specificity and their corresponding variances are combined to provide final estimates. The predictive distribution of the missing data is derived given the observed data and sampling iteratively from multinomial distribution and posterior distribution. Harel and Zhou conclude that the proposed estimators are better than the estimators of Begg and Greenes (Equations 3.2 and 3.3). However, there has been debate about the validity of this conclusion (Hanley *et al.*, 2007; Harel and Zhou, 2007). Subsequently, De Groot and colleagues identified computational errors in the work of Harel and Zhou (2006) which make it difficult to accurately draw conclusions from their work. Therefore, a separate comparison of the multiple imputation estimator and Begg and Greenes estimator was performed (De Groot *et al.*, 2011). The conclusion of this comparison is that both estimation methods yield similar results when the missing data mechanism is straightforward, but multiple imputation is recommended when the missing data mechanism is less straightforward or unknown.

3.2.2. NI approaches

If the decision to obtain disease verification depends on unrecorded factors related to disease, then the MAR assumption is not satisfied and the estimators discussed above could be biased. Zhou (1993) extended Begg and Greenes' method to allow a more general model for the verification process and derived the maximum likelihood estimators for the sensitivity and specificity of a diagnostic test and their corresponding variances. This approach does not assume D is MAR, but assumes that

$$\lambda_1 = \frac{P\big(V=1 \mid D=1, T=1\big)}{P\big(V=1 \mid D=0, T=1\big)}\ , \qquad \lambda_0 = \frac{P\big(V=0 \mid D=1, T=1\big)}{P\big(V=0 \mid D=0, T=1\big)}\ ,$$

are known. In other words, the ratio of the probability of selecting for verification a diseased patient with a given test result to that of selecting for verification a non-diseased patient with the same test result is known. In practice, however, λ_1 and λ_0 are not usually known and may be difficult to estimate. If $\lambda_1 = \lambda_0$, then Zhou's estimators reduce to those of Begg and Greenes.

Kosinski and Barnhart (2003) derive a region of all sensitivity and specificity values consistent with the observed data. This region is referred to as the test ignorance region. Recall that disease verification is not determined for u_1 test positives and u_0 test negatives. Of the u_1 test positives, let u_{1D} correspond to those truly diseased so there are $u_1 - u_{1D}$ test positives that are truly non-diseased. Similarly, let u_{0D} correspond to the truly diseased test negatives so there are $u_0 - u_{0D}$ test negatives that are truly non-diseased. If these values were known, then sensitivity (sens) and specificity (spec) can be estimated as

$$\text{sens} = \frac{s_1 + u_{1D}}{s_1 + u_{1D} + s_0 + u_{0D}}\ , \qquad \text{spec} = \frac{r_0 + u_0 - u_{0D}}{r_0 + u_0 - u_{0D} + r_1 + u_1 - u_{1D}}\ .$$

The test ignorance region is a plot of all sensitivity and specificity values resulting by considering all possible values of u_{1D} and u_{0D} in these equations.

An interactive web-based tool has been developed (Richardson and Petscavage (2010)) to implement the global sensitivity analysis of Kosinski and Barnhart. This tool is available at `http://uwmsk.org/gsa`. We illustrate this tool by using the coronary artery disease data summarized in Table 3. The region between the two curves in Figure 1 corresponds to the test ignorance region of all sensitivity and specificity values consistent with the observed data. The Begg and Greenes estimates (labeled MAR) fall in this region while the complete case or unadjusted estimates (labeled MCAR) fall outside the region and are therefore not compatible with the data.

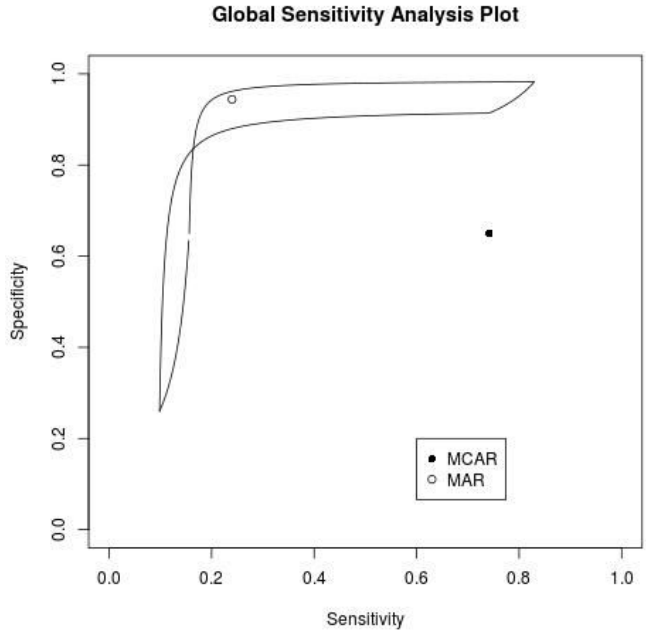

Figure 1: Global sensitivity analysis of the coronary artery disease data. MAR corresponds to Begg and Greenes estimates. MCAR corresponds to complete case estimates.

Baker (1995) and Kosinski and Barnhart (2003) propose likelihood-based regression approaches to deal with NI missingness when estimating the accuracy of a dichotomous test. These approaches require multiple diagnostic tests or covariates X. The approaches differ in how they factor the joint probability $P(V, T, D)$ as the product of conditional probabilities. Baker considered

$$P(V, T, D \mid X) \;=\; P(T \mid X)\, P(D \mid T, X)\, P(V \mid T, D, X)$$

while Kosinski and Barnhart considered

$$P(V, T, D \mid X) \;=\; P(D \mid X)\, P(T \mid D, X)\, P(V \mid T, D, X)\ .$$

The latter formulation is a product of the disease component $P(D \mid X)$, diagnostic test component $P(T \mid D, X)$, and missing data mechanism component $P(V \mid D, T, X)$. This formulation has the nice feature that sensitivity and specificity can be obtained directly from the diagnostic test component. Logistic regression models can be used to estimate parameters for each of the three components. When D is not included as a covariate in the missing data mechanism model, the missingness is MAR. Therefore, likelihood ratio, Wald, or Score tests can be used to test whether the MAR assumption is valid by testing whether the parameter is zero for D in the logistic regression model for $P(V \mid D, T, X)$. The expectation and maximization (EM) algorithm can be used to determine maximum likelihood estimates.

3.2.3. Bayesian approaches

Two Bayesian approaches have been developed to adjust for verification bias when estimating sensitivity and specificity of a binary diagnostic test. Both approaches allow for NI missingness. Martinez *et al.* (2006) describes an empirical Bayesian approach where Beta prior distributions are assumed for sensitivity, specificity, prevalence of disease, and the ratio of the probability of selecting for verification a diseased patient with a given test result to that of selecting for verification a non-diseased patient with the same test result is known (λ_1 and λ_0 considered by Zhou (1993)). Prior distributions for sensitivity and specificity are based on Begg and Greenes estimates of sensitivity and specificity and non-informative priors are used for the other parameters. The Gibbs sampling algorithm is used to estimate marginal posterior densities for all parameters.

Buzoianu and Kadane (2008) use the formulation $P(V,T,D)$ is equal to $P(D)\,P(T \mid D)\,P(V \mid T,D)$ considered by Kosinski and Barnhart (2003) to accommodate NI missingness. Similar to Kosinksi and Barnhart, logistic regression models can be used for each component. Prior distributions are used for the parameters in the logistic models.

4. CONTINUOUS TEST

Consider a continuous test T where higher values of T are more indicative of disease. The accuracy of a continuous diagnostic test is typically assessed using an ROC curve. An ROC curve is a plot of the true positive rate (TPR), sensitivity, versus the false positive rate (FPR), one minus the specificity, associated with all the dichotomous tests that can be formed by varying the cut point that defines a positive dichotomous test. When all subjects are verified, TPR and FPR can be estimated nonparametrically for a particular cutpoint c by using

$$\widehat{\mathrm{TPR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\, D_i}{\sum_{i=1}^{n} D_i}\ , \qquad \widehat{\mathrm{FPR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\,(1-D_i)}{\sum_{i=1}^{n}(1-D_i)}\ .$$

4.1. Impact of bias

Complete case estimators only use data from subjects who received disease verification. That is,

$$\widehat{\mathrm{TPR}}(c)_{CC} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\, V_i D_i}{\sum_{i=1}^{n} V_i D_i}\ , \qquad \widehat{\mathrm{FPR}}(c)_{CC} = \frac{\sum_{i=1}^{n} I(T_i \geq c)\, V_i (1-D_i)}{\sum_{i=1}^{n} V_i (1-D_i)}\ .$$

The complete case estimator yields unbiased estimates of the ROC curve and corresponding AUC when disease verification is MCAR. If the missing data mechanism is not MCAR, the complete case estimator can yield biased estimates of the ROC curve by overestimating $\text{TPR}(c)$ and $\text{FPR}(c)$ for each cutpoint c that results in operating points on the ROC curve that are biased upwards relative to the full data curve and thus underestimates the ROC curve and corresponding AUC. However, the complete case approach can also overestimate the ROC curve and AUC depending on the verification mechanism and accuracy of T (Alonzo and Pepe, 2005).

4.2. Bias correction—ROC curve

4.2.1. MAR approaches

Alonzo and Pepe (2005) proposed several bias-corrected estimators of TPR and FPR that assume disease status is MAR. Bias-corrected ROC curves are obtained by plotting bias-corrected estimators of TPR and FPR for all cutpoints. One approach for bias correction is to use full imputation (FI) over the distribution $P(D \mid T, X)$. That is, FI imputes $\rho = P(D \mid T, X)$ for all subjects in the study which results in the following estimators

$$\widehat{\text{TPR}}_{\text{FI}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ \widehat{\rho}_i}{\sum_{i=1}^{n} \widehat{\rho}_i}\ , \qquad \widehat{\text{FPR}}_{\text{FI}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ (1-\widehat{\rho}_i)}{\sum_{i=1}^{n} (1-\widehat{\rho}_i)}\ ,$$

where $\widehat{\rho}_i$ is an estimate of $P(D_i = 1 \mid T_i, X_i)$ that can obtained using, for example, logistic regression. By the MAR assumption, the disease model $P(D = 1 \mid T, X)$ can be estimated using the verification sample. When T and X are discrete and a saturated model is used, these estimators of TPR and FPR reduce to the Begg and Greenes (1983) bias-corrected estimators of sensitivity and specificity presented in the previous section.

Another approach for bias correction is to use mean score imputation (MSI) where the observed disease status is used for those in the verification sample and disease status is imputed for subjects not in the verification sample. That is,

$$\widehat{\text{TPR}}_{\text{MSI}}(c) \ = \ \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i\, D_i + (1-V_i)\, \widehat{\rho}_i \right\}}{\sum_{i=1}^{n} \left\{ V_i\, D_i + (1-V_i)\, \widehat{\rho}_i \right\}}\ ,$$

$$\widehat{\text{FPR}}_{\text{MSI}}(c) \ = \ \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i\, (1-D_i) + (1-V_i)\, (1-\widehat{\rho}_i) \right\}}{\sum_{i=1}^{n} \left\{ V_i\, (1-D_i) + (1-V_i)\, (1-\widehat{\rho}_i) \right\}}\ .$$

Again, the MAR assumption implies that data from the verification sample can be used to obtain valid estimates of ρ_i.

Alonzo and Pepe (2005) also propose the following inverse probability weighting (IPW) estimators (Horvitz and Thompson, 1952) that weight each observation in the verification sample by the inverse of the sampling fraction (*i.e.* probability the subject was selected for verification)

$$\widehat{\text{TPR}}_{\text{IPW}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ V_i\, D_i \,/\, \widehat{\pi}_i}{\sum_{i=1}^{n} V_i\, D_i \,/\, \widehat{\pi}_i} ,$$

$$\widehat{\text{FPR}}_{\text{IPW}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\ V_i(1-D_i) \,/\, \widehat{\pi}_i}{\sum_{i=1}^{n} V_i(1-D_i) \,/\, \widehat{\pi}_i} ,$$

where $\widehat{\pi}_i = P(V_i = 1 \,|\, T_i, X_i)$ may be known or may need to be estimated depending on the design of the study. The IPW estimators are similar to the CC estimators in that they use the observed disease status for the verification sample. Unlike the CC, however, they correct for the biased sampling by weighting the observed value by the probability the subject was verified.

The following doubly robust (DR) estimators have also been proposed:

$$\widehat{\text{TPR}}_{\text{DR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i D_i/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, \widehat{\rho}_i/\widehat{\pi}_i \right\}}{\sum_{i=1}^{n} \left\{ V_i D_i/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, \widehat{\rho}_i/\widehat{\pi}_i \right\}} ,$$

$$\widehat{\text{FPR}}_{\text{DR}}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c) \left\{ V_i(1-D_i)/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, (1-\widehat{\rho}_i)/\widehat{\pi}_i \right\}}{\sum_{i=1}^{n} \left\{ V_i(1-D_i)/\widehat{\pi}_i - (V_i - \widehat{\pi}_i)\, (1-\widehat{\rho}_i)/\widehat{\pi}_i \right\}} .$$

These estimators are referred to as doubly robust because they are consistent if either π_i or ρ_i is estimated consistently. That is, the verification model or disease model can be incorrectly specified and consistency is still guaranteed. These estimators have also been referred to as semiparametric because they require parametric conditional mean models to be specified for the disease model $P(D \,|\, T, X)$ and for the verification model $P(V \,|\, T, X)$ but are non-parametric with respect to the joint distribution of the data $P(D, T, X)$.

Alonzo and Pepe (2005) illustrated that misspecifying the verification model yields biased IPW estimates of the ROC curve and misspecifying the disease model results in biased FI and MSI. Furthermore, they showed the DR estimator of the ROC curve is unbiased if either the model for verification or the model for disease is correctly specified. Thus, they recommend the DR approach is used in practice.

The AUC can be estimated empirically for each of the bias-corrected ROC curves described above by using the Trapezoidal Rule (Bamber, 1975). Closed-form expressions for the AUC corresponding to the IPW and DR ROC estimators have been obtained as well as variance expressions (He *et al.*, 2009).

4.2.2. NI approaches

Rotnitzky *et al.* (2006) describe a DR estimator of the AUC. They note that AUC is identified under the untestable assumption

$$\log\left\{\frac{P(V=0\,|\,T,X,V)}{P(V=1\,|\,T,X,V)}\right\} \;=\; h(T,V)+q(T,V)\,X\;, \tag{4.1}$$

where $q(T,V)$ is an arbitrary specified function and $h(T,V)$ is an arbitrary unknown function. $q(T,V)$=0 for all T and V corresponds to the MAR assumption while $q(T,V)\neq 0$ corresponds to NI missingness. Fluss *et al.* (2009) extend the approach of Rotnitzky *et al.* (2006) to obtain a DR estimate of TPR and FPR and, thus, the empirical ROC curve that allows for NI missingness. They recommend performing a sensitivity analysis by repeating the estimation of TPR and FPR under a variety of reasonable choices for the selection bias function q. Conversely, Liu and Zhou (2010) use the likelihood approach to estimate a nonignorable parameter and obtain DR estimates of the ROC curve and AUC. They assume the disease verification model

$$P\big(V_i=1\,|\,D_i,T_i,X_i\big) \;=\; \frac{\exp(x)}{1+\exp(x)}\left\{h(T_i,X_i;\beta)+\alpha\,D_i\right\}\;,$$

where α is the NI parameter and $h(T_i,X_i;\beta)=\beta_0+\beta_1T_i+\beta_2X_i$. Since the nonignorable parameter cannot be tested nonparametrically, Liu and Zhou recommend that scientific knowledge is used to construct an appropriate disease verification model.

4.3. Covariate-adjusted ROC curves

The accuracy of a diagnostic test can be affected by factors such as disease severity, age, and gender. ROC curves have been adjusted for age in the assessment, for example, of the accuracy of fingerstick postprandial blood glucose measurements to discriminate between healthy and diseased subjects in the presence of verification bias (Fluss *et al.*, 2012).

Page and Rotnitzky (2009) discuss a parametric model for estimating the covariate-specific ROC curve in the presence of verification bias. They make the assumption that the ROC curve has an underlying binormal distribution and disease verification has NI missingness. Liu and Zhou (2011) discuss a likelihood approach to estimate the covariate-specific ROC curve in the presence of verification bias. Disease verification is assumed to be MAR and diagnostic test results are modeled using a location-scale model. Weighted estimating equations are used to estimate the parameters in the location-scale model. DR, IPW, and imputation approaches are compared for the estimation. Liu and Zhou conclude that

the DR estimator performed best in their simulation studies and their method is sensitive to the location-scale model assumption.

Fluss *et al.* (2012) develop a DR method for estimating the ROC curve adjusted for covariates for a NI missing data mechanism. Using the approach of Pepe (1998), they model the diagnostic test values distribution as a function of disease status and covariate values using a semi-parametric location-scale model. Since the proposed approach relies on the untestable specification of $q(T, V)$ (see Equation 4.1), the authors recommend a sensitivity analysis is performed to examine the sensitivity of the estimated ROC curve to the specified form of $q(T, V)$.

5. DISCUSSION

This paper highlights methods available for estimating the accuracy of dichotomous and continuous diagnostic tests in the presence of verification bias. More recently, this bias has also been referred to as partial verification bias so as not to be confused with differential disease verification in which a subset of study subjects have a different reference standard to determine disease status (Whiting, 2004).

As investigators design future studies of test accuracy, it is important to record all factors that may affect the decision to offer and receive disease verification. In cases where all factors are captured, then the MAR assumption will likely be satisfied and bias-correction methods that rely on this assumption can be used. When all factors that impact disease verification are not collected, it is preferred to use bias-correction methods that allow for NI missingness.

The focus of this paper is on the estimation of the sensitivity and specificity of a single dichotomous test and the ROC curve and AUC for a single continuous test in the presence of verification bias. Bias correction methods are also available for diagnostic tests measured on an ordinal scale (Gray *et al.*, 1984; Hunink *et al.*, 1990; Baker, 1995; Toledano and Gatsonis, 1996; Rodenberg and Zhou, 2000), such as a radiologist's interpretations of images to quantify the suspicion of cancer. In addition, methods have been developed to estimate the difference between two diagnostic tests in regards to bias-corrected sensitivity and specificity. Assuming disease verification is MAR, Zhou (1998) and Roldán Nofuentes and Luna del Castillo (2008) provide estimators for the difference in bias-corrected sensitivity and specificity.

This paper considers the setting when there are only two disease states (diseased and non-diseased). In some settings there can be more than two disease states. For example, Alzheimer's Disease dementia can be classified into more

than two categories. Chi and Zhou (2008) propose a non-parametric likelihood-based approach to construct the empirical ROC surface (extension of ROC curve to more than two disease states) and estimate the volume under the ROC surface in the presence of verification bias for ordinal diagnostic tests. Future work is needed to develop bias correction methods for estimating the ROC surface and volume under the ROC surface for continuous diagnostic tests.

The bias correction methods described in this paper, especially for continuous tests, would benefit from the development and distribution of code to apply the methods in practice. Increasing the availability of these methods in standard statistical packages would likely increase the use of the methods.

REFERENCES

Alonzo, T. A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias, *Journal of the Royal Statistical Society, Ser.* C, **54**, 173–190.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph, *Journal of Mathematical Psychology*, **12**, 387–415.

Baker, S. G. (1995). Evaluating multiple diagnostic tests with partial verification, *Biometrics*, **51**, 330–337.

Begg, C. B. and Greenes, R. A. (1983). Assessment of diagnostic tests when disease is subject to selection bias, *Biometrics*, **39**, 207–216.

Buzoianu, M. and Kadane, J. B. (2008). Adjusting for verification bias in diagnostic test evaluation: a Bayesian approach, *Statistics in Medicine*, **27**, 2453–2473.

Chi, Y.-Y. and Zhou, X.-H. (2008). Receiver operating characteristic surfaces in the presence of verification bias, *Journal of the Royal Statistical Society, Ser.* C, **57**, 1–23.

De Groot, J. A. H.; Janssen, K. J. M.; Kristel, J. M.; Zwinderman, A. H.; Bossuyt, P. M. M.; Reitsma, J. B. and Moons, K. G. M. (2011). Correcting for partial verification bias: a comparison of methods, *Annals of Epidemiology*, **21**, 139–148.

De Groot, J.A.H.; Janssen, K.J.M.; Zwinderman, A.H.; Moons, K.G.M. and Reitsma, J. B. (2008). Multiple imputation to correct for partial verification bias revisited, *Statistics in Medicine*, **27**, 5880–5889.

Fluss, R.; Reiser, B. and Faraggi, D. (2012). Adjusting ROC curves for covariates in the presence of verification bias, *Journal of Statistical Planning and Inference*, **142**, 1–11.

Fluss, R.; Reiser, B.; Faraggi, D. and Rotnitzky, A. (2009). Estimation of the ROC curve under verification bias, *Biometrical Journal*, **51**, 475–490.

Gray, R.; Begg, C. and Greenes, R. (1984). Construction of receiver operating characteristic curves when disease verification is subject to selection bias, *Medical Decision Making*, **4**, 151–164.

Hanley, J. A.; Dendukuri, N. and Begg, C. B. (2007). Multiple imputation for correcting verification bias by Ofer Harel and Xiao-Hua Zhou, *Statistics in Medicine*, **26**, 3046–3047.

Harel O. and Zhou X.-H. (2006). Multiple imputation for correcting verification bias, *Statistics in Medicine*, **25**, 3769–3786.

Harel O. and Zhou X.-H. (2007). Rejoinder to multiple imputation for correcting verification bias, *Statistics in Medicine*, **26**, 3047–3050.

He, H.; Lyness, J. M. and McDermott, M. P. (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias, *Statistics in Medicine*, **28**, 361–376.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe, *Journal of the American Statistical Association*, **47**, 663–685.

Hunink, M. G. M.; Richardson, D. K.; Doubilet, P. M. and Begg, C. B. (1990). Testing for fetal pulmonary maturity ROC analysis involving covariates, verification bias, and combination testing, *Medical Decision Making*, **10**, 201–211.

Iglesias-Garriz, I.; Rodríguez, M. A.; García-Porrero, E.; Ereño, F.; Garrote, C. and Suarez, G. (2005). Emergency nontraumatic chest pain: use of stress echocardiography to detect significant coronary artery stenosis, *Journal of the American Society of Echocardiography*, **18**, 1181–1186.

Kosinski, A. S. and Barnhart, H. X. (2003). A global sensitivity analysis of performance of a medical diagnostic test when verification bias is present, *Statistics in Medicine*, **22**, 2711–2721.

Kosinski, A. S. and Barnhart, H. X. (2003). Accounting for nonignorable verification bias in assessment of diagnostic tests, *Biometrics*, **59**, 163–171.

Little, R. J. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*, Wiley, New York.

Liu, D. and Zhou, X.-H. (2010). A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach, *Biometrics*, **66**, 1119–1128.

Liu, D. and Zhou, X.-H. (2011). Semiparametric Estimation of the Covariate-Specific ROC Curve in Presence of Ignorable Verification Bias, *Biometrics*, **67**, 906–916.

Martinez, E. Z.; Alberto Achcar, J. and Louzada-Neto, F. (2006). Estimators of sensitivity and specificity in the presence of verification bias: A Bayesian approach, *Computational Statistics and Data Analysis*, **51**, 601–611.

Page, J. H. and Rotnitzky, A. (2009). Estimation of the disease-specific diagnostic marker distribution under verification bias, *Computational Statistics and Data Analysis*, **53**, 707–717.

Pepe, M. S. (1998). Regression analysis of ROC curves, *Biometrics*, **54**, 124–135.

Ransohoff, D. F. and Feinstein, A. R. (1978). Problems of spectrum and bias in evaluating the efficacy of diagnostic tests, *New England Journal of Medicine*, **299**, 926–930.

Richardson, M. L. and Petscavage, J. M. (2010). An interactive web-based tool for detecting verification (work-up) bias in studies of the efficacy of diagnostic imaging, *Academic Radiology*, **17**, 1580–1583.

Rodenberg, C. and Zhou, X.-H. (2000). ROC curve estimation when covariates affect the verification process, *Biometrics*, **56**, 131–136.

Roldán Nofuentes, J. A. and Luna del Castillo, J. D. (2008). EM algorithm for comparing two binary diagnostic tests when not all the patients are verified, *Journal of Statistical Computation and Simulation*, **78**, 19–35.

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.

Toledano, A. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data, *Statistics in Medicine*, **15**, 1807–1826.

Thompson, I. M.; Ankerst, D. P.; Chi, C.; Lucia, M. S.; Goodman, P. J.; Crowley, J. J.; Parnes, H. L. and Coltman Jr, C. A. (2005). Operating characteristics of prostate-specific antigen in men with an initial PSA level of 3.0 ng/ml or lower, *The Journal of the American Medical Association*, **294**, 66–70.

Whiting, P.; Rutjes, A. W. S.; Reitsma, J. B.; Glas, A. S.; Bossuyt, P. M. M. and Kleijnen, J. (2004). Sources of Variaton and Bias in Studies of Diagnostic Accuracy: A Systematic Review, *Annals of Internal Medicine*, **140**, 189–202.

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

Zhou, X.-H. (1994). Effect of verification bias on positive and negative predictive values, *Statistics in Medicine*, **13**, 1737–1745.

Zhou, X.-H. (1998). Comparing accuracies of two screening tests in a two-phase study for dementia, *Journal of the Royal Statistical Society, Ser.* C, **47**, 135–147.