Published in final edited form as:
Biom J. 2009 June ; 51(3): 475–490. doi:10.1002/bimj.200800128.

Estimation of the ROC Curve under Verification Bias

RONEN FLUSS[1]
Department of Health Services Research, Ministry of Health, Jerusalem, Israel rfluss@gmail.com

BENJAMIN REISER and **DAVID FARAGGI**
Department of Statistics, University of Haifa, Haifa, Israel

ANDREA ROTNITZKY
Department of Economics, Di Tella University, Buenos Aires, Argentina

Department of Biostatistics, Harvard School of Public Health, Boston, MA

Summary

The ROC (Receiver Operating Characteristic) curve is the most commonly used statistical tool for describing the discriminatory accuracy of a diagnostic test. Classical estimation of the ROC curve relies on data from a simple random sample from the target population. In practice, estimation is often complicated due to not all subjects undergoing a definitive assessment of disease status (verification). Estimation of the ROC curve based on data only from subjects with verified disease status may be badly biased. In this work we investigate the properties of the doubly robust (DR) method for estimating the ROC curve under verification bias originally developed by Rotnitzky *et al.* (2006) for estimating the area under the ROC curve. The DR method can be applied for continuous scaled tests and allows for a non ignorable process of selection to verification. We develop the estimator's asymptotic distribution and examine its finite sample properties via a simulation study. We exemplify the DR procedure for estimation of ROC curves with data collected on patients undergoing electron beam computer tomography, a diagnostic test for calcification of the arteries.

Keywords

Diagnostic test; Nonignorable; Semiparametric model; Sensitivity analysis; Sensitivity; Specificity

1 Introduction

The ROC (Receiver Operating Characteristic) curve is the most commonly used summary measure for describing the discriminatory accuracy of a diagnostic test in distinguishing between diseased and healthy individuals. Ideally, the estimation of the ROC curve relies on a random sample from the target population comprised of healthy and diseased subjects. However, often in practice, not all subjects undergo the definitive assessment of disease since the verification procedure is expensive, invasive or both. This results in some individuals having missing information on their disease status. The decision to send a patient to verification is often based on the test result and other patient characteristics. As noted by many authors (Begg and Greenes 1983; Zhou, 1994, 1998a) estimators of the ROC curve and other summary measures of test performance based on data from patients with verified disease status only may be badly biased. This bias is usually referred to as verification bias.

[1] To whom correspondence should be addressed.

The methods developed in the literature for correcting verification bias are usually based on the following assumptions that limit their usefulness in applications. First, most currently available methods assume that the diagnostic test's scale is ordinal (Begg and Greenes, 1983; Gray *et al.*, 1984; Baker, 1995; Toledano and Gatsonis 1996; Zhou, 1996, 1998a,b,c; Rodenberg and Zhou, 2000). Yet many important emerging tests are measured on a continuous scale. Second, with the exception of few articles which only deal with dichotomous diagnostic tests (e.g. Baker, 1995; Zhou, 1993, 1994; Kosinski and Barnhart, 2003), currently available methods assume that the decision to send a patient to verification is conditionally independent of the true disease status of the patient given the test results and possibly other observed covariates (see Alonzo and Pepe, 2005, for continuous scales), or equivalently, that the missing disease status is missing at random (MAR) (Rubin,1976). However, usually the doctor's decision to send a patient to verification will be based on his/her detailed information on the patient's health, which can hardly ever be accurately summarized by the test result and other measured covariates. The aforementioned methods can yield biased estimates in this case because the assumption of MAR data no longer holds since non-response and true disease status are dependent even after adjusting for measured variables. In such a case, the missing process on the disease status is known as non-ignorable.

As an example, consider a study run by the Nuclear Imaging Group at Cedars Sinai Medical Center discussed by Rotnitzky *et al.* (2006). A diagnostic test for coronary artery disease, electron beam computed tomography (EBCT) (Braun *et al.*, 1996), was performed on 5130 males, of which only 379 of these had disease status checked with the more expensive Dual Isotope Myocardial Perfusion Single Photon Emission Computed Tomography (SPECT) test. Of the verified subjects only 28 were found to be diseased. The EBCT distribution of non-verified subjects is markedly skewed to the right (when compared to the verified subjects) indicating that subjects with lower values of the marker are less likely than others to be verified. It is indeed very likely that doctors based their decision to send a patient to have a SPECT test not just on the values of the EBCT marker, but on other clinical variables not available for data analysis.

Rotnitzky *et al.* (2006) discuss an alternative estimation approach in the context of the estimation of the Area under the ROC curve (AUC). Their approach can be used with either continuous or ordinal markers and is especially suitable for conducting sensitivity analysis to different degrees of residual association between selection to verification and true disease status after adjusting for the markers and other measured explanatory variables. In Section 2 we review Rotnitzky *et al.* approach for correcting for verification bias, extend this method to the estimation of the ROC curve and obtain the asymptotic properties of our estimator. In Section 3 we discuss a few extensions. In Section 4 we report the results of a simulation study. In Section 5 we apply the proposed methods to the EBCT data and conclude with a discussion in Section 6.

2 Adjusting for Verification Bias

Suppose that on each subject i of a random sample of n patients, we measure a diagnostic test result Y_i which can be continuous or ordinal, and a covariate vector V_i of health and demographic variables. Let X_i be the indicator of true disease status of patient i, $X_i = 1$ if patient i has the disease of concern, and $X_i = 0$ otherwise. Suppose that X_i is missing in a subset of the study participants because not all subjects undergo the definitive assessment of disease. Let $R_i = 1$ if patient i is sent to verification (i.e. if X_i is observed) and $R_i = 0$ otherwise. The vectors (X_i, Y_i, V_i, R_i), $i = 1, ..., n$ are independent identically distributed (i.i.d) copies of a random vector (X, Y, V, R). Under our model the observed data are

comprised of n i.i.d copies O_i of the random vector $O = (R, c(R, X, Y, V))$ where $c(1, X, Y, V) = (X, Y, V)$ and $c(0, X, Y, V) = (Y, V)$.

Suppose that a subject i were to be classified as diseased if $T_i = I(Y_i > c)$ is equal to 1 and as non-diseased if $T_i = 0$, for a given constant c. The sensitivity, $Se(c) = \Pr(Y > c|X = 1)$, and specificity, $Sp(c) = \Pr(Y < c|X = 0)$, would then be the probability of diagnosing a diseased and a healthy person correctly, respectively. For notational convenience, in what follows the c will on occasions be dropped. The theoretical ROC curve is the plot of sensitivity against 1-specificity for all possible values of c. Consequently estimation of the ROC curve follows immediately from estimation of the $\theta(c) = (Se(c), Sp(c))$ pairs. Supposing that all patients are verified (i.e. $\forall i\ R_i = 1$), the empirical estimator

$$\widehat{\theta} = \left(\widehat{S}_e, \widehat{S}_p\right)' = \left(\frac{\sum_{i=1}^{n} X_i T_i}{\sum_{i=1}^{n} X_i}, \quad \frac{\sum_{i=1}^{n} (1 - X_i)(1 - T_i)}{\sum_{i=1}^{n} (1 - X_i)} \right)' \quad (2.1)$$

is consistent for $\theta(c)$. If not all subjects undergo verification, i.e. if R_i is not 1 for all i, but the process of selection to verification is completely random, i.e. if R_i and X_i are independent, then $\widehat{\theta}$ calculated from verified subjects only remains consistent for $\theta(c)$. However, if R_i and X_i are correlated, then this naive estimator will lead to inconsistent estimates of θ.

The key issue that complicates the analysis is that both $Se(c)$ and $Sp(c)$ are unidentified from the observed data O_i, $i = 1, ..., n$. That is, there exist many curves $Se(c)$ and $Sp(c)$, $c \in \mathbf{R}$, compatible with the distribution of O_i, so that even in the ideal situation in which the distribution of the observed data were perfectly known, there would remain uncertainty about the true ROC curve. To identify the ROC curve and associated summary measures one needs to make untestable assumptions, i.e. assumptions that cannot be rejected by any statistical test no matter how large n is. Rotnitzky *et al.* (2006) argued that the area under the ROC curve is identified under the following untestable assumption,

$$\log\left\{\frac{\Pr(R=0|Y,X,V)}{\Pr(R=1|Y,X,V)}\right\} = h(Y,V) + q(Y,V)\,X \quad (2.2)$$

where $q(Y, V)$ is an arbitrary specified (i.e. known) function and $h(Y, V)$ is an arbitrary but unknown function. Indeed, this assumption suffices also to identify $\theta(c)$ for any fixed c.

Application of Bayes rule shows that (2.2) holds for some $h(Y, V)$ if and only if

$$\Pr(X=x|R=0,Y,V) = \Pr(X=x|R=1,Y,V)\,\frac{\exp(q(Y,V)\,x)}{c(Y,V)} \quad (2.3)$$

where $c(Y, V) = E\{\exp(q(Y, V)\,X)\,|R = 1, Y, V\}$ (Rotnitzky *et al.*, 2006). The function $q(Y, V)$ measures the degree of the residual association between R and X within levels of Y and V. For example, the choice $q(Y, V) = -1$ indicates that the residual association is constant within levels of (Y, V) and such that among subjects with identical values of (Y, V), the odds of being sent to verification is $\exp(1) = 2.71$ times larger for diseased subjects than for non-diseased subjects. The choice $q(Y, V) = 0$ for all V and Y corresponds to the assumption that the missing disease status is missing at random. In general, when $q(Y, V) \neq 0$, the selection process is said to be non-ignorable.

We refer to the model that assumes just (2.2) with specified $q(Y, V)$ and $h(Y, V)$ unknown (or equivalently equation (2.3) with specified $q(Y, V)$) as model $\mathscr{A}(q)$. Arguing as in Scharfstein *et al.* (1999), it can be shown that imposing a given choice of q does not restrict the observed data distribution. That is, model $\mathscr{A}(q)$ is a non-parametric model for the distribution of O. Since all values of the selection bias parameter determine the same model for the observed data distribution, the selection bias parameter is untestable, because any q will perfectly fit the observed data. Following Robins *et al.* (2000) we therefore recommend conducting inference about $\theta(c)$ by repeating the estimation under different plausible choices for the selection bias function q as a form of sensitivity analysis to different degrees of selection bias. This raises the question of how to choose the selection bias functions in practice. We suggest that one chooses a collection of simple selection bias functions indexed by one or two parameters β that are to be varied in a sensitivity analysis, with values of the parameters equal to zero corresponding to the function $q = 0$. For example, the choice $q(Y, V) = \beta$ is tantamount to assuming that the odds of verification of diseased vs non-diseased subjects is constant across all values of the marker and the auxiliary V and equal to $\exp(-\beta)$. This is the choice used in our simulation study in Section 4. One could consider the choice $q(Y, V) = \beta_0 + \beta_1 Y$ if the investigator believed that the odds of verification of diseased vs non-diseased subjects is modified by the values of Y, this modification being monotone in Y.

The identity (2.3) implies that

$$E(X|Y, V, R) = E(X|Y, V, R=1)\left\{R + \frac{(1-R)\exp\{q(Y,V)\}}{1 - E(X|Y, V, R=1)\{1 - \exp\{q(Y,V)\}\}}\right\}.$$

Consequently, writing $Se(c)$ as $E\{E(X/Y, V, R)\, T\} / E\{E(X/Y, V, R)\}$ and $Sp(c)$ as $E\{[1 - E(X/Y, V, R)](1 - T)\} / E\{[1 - E(X/Y, V, R)]\}$ we obtain that under model $\mathscr{A}(q)$

$$Se(c) = \frac{E\left[E(X|Y, V, R=1)\left\{R + \frac{(1-R)\exp\{q(Y,V)\}}{1-E(X|Y,V,R=1)\{1-\exp\{q(Y,V)\}\}}\right\} T\right]}{E\left[E(X|Y, V, R=1)\left\{R + \frac{(1-R)\exp\{q(Y,V)\}}{1-E(X|Y,V,R=1)\{1-\exp\{q(Y,V)\}\}}\right\}\right]} \quad (2.4)$$

and

$$Sp(c) = \frac{E\left[\left[1 - E(X|Y, V, R=1)\left\{R + \frac{(1-R)\exp\{q(Y,V)\}}{1-E(X|Y,V,R=1)\{1-\exp\{q(Y,V)\}\}}\right\}\right](1-T)\right]}{E\left[\left[1 - E(X|Y, V, R=1)\left\{R + \frac{(1-R)\exp\{q(Y,V)\}}{1-E(X|Y,V,R=1)\{1-\exp\{q(Y,V)\}\}}\right\}\right]\right]}. \quad (2.5)$$

Likewise, letting $\pi(Y, V)$ denote the quantity $(1 + \exp\{h(Y, V) + q(Y, V) X\})^{-1}$, the identity (2.2) implies that $P(R = 1|X, Y, V) = \pi(Y, V)$. Consequently, under modelg $\mathscr{A}(q)$, $Se(c)$ and $Sp(c)$ can be expressed also as

$$Se(c) = \frac{E\left\{\frac{R}{\pi(Y,V)} XT\right\}}{E\left\{\frac{R}{\pi(Y,V)} X\right\}} \quad \text{and} \quad Sp(c) = \frac{E\left\{\frac{R}{\pi(Y,V)}(1-X)(1-T)\right\}}{E\left\{\frac{R}{\pi(Y,V)}(1-X)\right\}}. \quad (2.6)$$

Model $\mathscr{A}(q)$ implicitly assumes that all subjects have a positive probability of being selected for verification regardless of the values of their test results and covariates. See Rotnitzky *et al.* (2006) for a discussion of the possibility of relaxing this condition. For technical reasons related to the finiteness of the variance of the estimators that we will propose later, we further need to assume that

$$\Pr(R=1|X, Y, V) \geq \sigma>0 \qquad \text{with probability} \quad 1. \quad (2.7)$$

Unfortunately, though sufficient for identification, model $\mathscr{A}(q)$ is insufficient for estimation of $\theta(c)$ due to the curse of dimensionality. Specifically, expressions (2.4), (2.5) and (2.6) imply that if one were to estimate $\theta(c)$ under model $\mathscr{A}(q)$, one would need to non-parametrically estimate either the function $h(Y, V)$ or the conditional expectation $E(X|Y, V, R = 1)$ using smoothing techniques, a practically unfeasible task when the dimension of V is large. Nevertheless, the expressions (2.4), (2.5) and (2.6) suggest two alternative dimension reducing strategies for estimating $\theta(c)$.

The first option is to assume that the unknown function $h(Y, V)$ follows a parametric model

$$h(Y, V) = h\left(Y, V; \gamma^0\right) \quad (2.8)$$

where $h(Y, V; \gamma)$ is smooth in γ and γ^0 is an unknown $p_h \times 1$ finite dimensional parameter. Assumption (2.8) and assumption (2.2) with specified $q(Y, V)$ define a semiparametric model for the observed data which for ease of reference we refer to as $\mathscr{B}(q)$. Note that this model imposes just a, possibly non-linear, logistic regression model on the selection probabilities with offset $q(Y, V)$ X. Expressions (2.6) imply that one can obtain a consistent and asymptotically normal estimator estimator of $\theta(c)$, throughout referred to as inverse probability weighted (IPW) estimator, under model $B(q)$ by replacing X_i in (2.1) with $R_i\pi(Y_i, V_i; \widehat{\gamma})^{-1} X_i$, where $\pi(Y, V; \gamma) = (1 + \exp\{h(Y, V; \gamma) + q(Y, V) X\})^{-1}$ and $\widehat{\gamma}$ is a consistent estimator of γ^0. Unfortunately, even though under model $\mathscr{B}(q)$, R follows a logistic regression on Y, X and V, we cannot find $\widehat{\gamma}$ from a standard logistic regression fit with offset because the offset $q(Y, V)$ X is not observed when $R = 0$. Instead, following Rotnitzky *et al.* (2006) we can compute $\widehat{\gamma}$ as the solution to

$$\Sigma_{i=1}^{n} B(\gamma; u)_i = 0 \quad (2.9)$$

where $B(\gamma; u) = \left\{R\pi(Y, V; \gamma)^{-1} - 1\right\} u(Y, V)$ and $u(Y, V)$ is an arbitrary, user specified, column vector function of the same dimension as γ. The choice of u impacts the efficiency with which we estimate γ. Using the theoretical results in Rotnitzky and Robins (1997), it can be shown that if $u(Y, V)$ is equal to

$$u_{opt}(Y, V) = \frac{\{\partial h(Y, V; \gamma)/\partial\gamma\}|_{\gamma=\gamma^0} E\{\exp[q(Y, V) X]|R=1, Y, V\}}{E\{\exp[q(Y, V) X]\pi^{-1}(Y, V; \gamma^0)|R=1, Y, V\}} \quad (2.10)$$

then the solution $\widehat{\gamma}_{opt}$ of (2.9) has asymptotic variance equal to the semiparametric variance bound for estimators of γ under model $\mathscr{B}(q)$. Note, in particular, that for the choice $q(Y, V) = 0$, $B(\gamma; u)$ reduces to $\{R - \pi(Y, V; \gamma)\}\{\partial h(Y, V; \gamma)/\partial\gamma\}|_{\gamma=\gamma^0}$ which, at γ^0, is equal to the score, i.e. the derivative of the log-likelihood for γ, in the logistic regression model logit $P(R = 1|Y, V) = -h(Y, V; \gamma)$.

A second option is to assume that $E(X|Y, V, R = 1) = \Pr(X = 1|R = 1, Y, V)$ follows a parametric model

$$\log\left\{\frac{\Pr(X=1|R=1, Y, V)}{\Pr(X=0|R=1, Y, V)}\right\} = m\left(Y, V; \mu^0\right) \quad (2.11)$$

where $m(Y, V; \mu)$ is a known function, smooth in μ, and μ^0 is an unknown $p_m \times 1$ parameter vector. Assumption (2.11) is a model for the disease probabilities among the verified subjects. The model defined by (2.3) with specified $q(Y, V)$ and (2.11) is a semiparametric model for the law of the observed data which we will denote with $\mathscr{C}(q)$. The Semi-Parametric Maximum Likelihood (SPML) estimator of $\theta(c)$ under model $\mathscr{C}(q)$ is obtained by replacing each X_i in (2.1) with $P(R_i, Y_i, V_i; \widehat{\mu})$, where

$$\begin{aligned} & P(R, Y, V; \mu) \\ = \quad & \Pr(X=1|R=1, Y, V; \mu)\left\{R+\frac{(1-R)\exp\{q(Y,V)\}}{1-\Pr(X=1|R=1,Y,V;\mu)\{1-\exp\{q(Y,V)\}\}}\right\} \end{aligned}$$

with $\Pr(X = 1|R = 1, Y, V; \mu) = \exp\{m(Y, V; \mu)\} / [1 + \exp\{m(Y, V; \mu)\}]$, and $\widehat{\mu}$ solves the score equations

$$\Sigma_{i=1}^{n} H(\mu)_i = 0$$

with $H(\mu) = R\frac{\partial m(Y, V; \mu)}{\partial \mu}\left\{X - [1+\exp\{-m(Y, V; \mu)\}]^{-1}\right\}$.

Neither option is entirely satisfactory because option 1 results in inconsistent estimation if model $\mathscr{B}(q)$ is incorrect and option 2 results in inconsistent estimation if model $\mathscr{C}(q)$ is misspecified. There is yet a third option which provides an estimator consistent for $\theta(c)$ under either model $\mathscr{B}(q)$ or model $\mathscr{C}(q)$ but not necessarily both. Specifically, define

$$X_{DR}(\gamma, \mu) = P(R, Y, V; \mu) + Q(\gamma, \mu)$$

where

$$Q(\gamma, \mu) = R[\{X - P(1, Y, V; \mu)\} + \exp\{h(Y, V; \gamma) + q(Y, V) X\}\{X - P(0, Y, V; \mu)\}].$$

Let $\widehat{\theta}_{DR}$ be the estimator of θ obtained by replacing each X_i in (2.1) with $X_{DR}(\widehat{\gamma}, \widehat{\mu})_i$. That is,

$$\widehat{\theta}_{DR} = \left(\frac{\sum_i X_{DR}(\widehat{\gamma}, \widehat{\mu})_i T_i}{\sum_i X_{DR}(\widehat{\gamma}, \widehat{\mu})_i}, \frac{\sum_i (1 - X_{DR}(\widehat{\gamma}, \widehat{\mu})_i)(1 - T_i)}{\sum_i (1 - X_{DR}(\widehat{\gamma}, \widehat{\mu})_i)}\right)'.$$

The estimator $\widehat{\theta}_{DR}$ is said to be double-robust because, as stated in Theorem 1 below, it is consistent and asymptotically normal for θ so long as model $\mathscr{A}(q)$, condition (2.7) and one of the models (2.11) or (2.8) holds, but not necessarily both. The key to the consistency of $\widehat{\theta}_{DR}$ lies in the fact that if model $\mathscr{A}(q)$ holds then, if model (2.11) additionally holds,

$$E\left\{X_{DR}\left(\gamma, \mu^0\right)|Y, V\right\} = E(X|Y, V) \quad \text{for any} \quad \gamma$$

or if model (2.8) additionally holds,

$$E\left\{X_{DR}\left(\gamma^0,\mu\right)|Y,V\right\}=E\left(X|Y,V\right) \quad \text{for any} \quad \mu \ .$$

In contrast to the first two estimators IPW and SPML, the double-robust estimator $\widehat{\theta}_{DR}$ gives the analyst two chances, instead of only one, to get nearly correct inference. Of course, there can be an efficiency price to using the DR estimator. If the disease regression model (2.11) is correct both the DR estimator and the SPML estimator will be consistent but the DR estimator will generally have larger (asymptotic) variance. Clearly, the efficiency gains over $\widehat{\theta}_{DR}$ are obtained at the cost of potential severe bias.

An interesting remark is that, invoking the theoretical results of Rotnitzky and Robins, 1997, it can be shown that if we use $u_{opt}(Y, V)$ to compute $\widehat{\gamma}$ then the $\widehat{\theta}_{DR}$ that uses this $\widehat{\gamma}$ is locally semiparametric efficient under model $\mathscr{B}(q)$ at the local model $\mathscr{C}(q)$. This means that, if both models $\mathscr{B}(q)$ and $\mathscr{C}(q)$ are correct, $\widehat{\theta}_{DR}$ has asymptotic variance that attains the semiparametric variance bound for estimators bof $\theta(c)$ under model $\mathscr{B}(q)$. This $\widehat{\theta}_{DR}$ is unfeasible because to compute $u_{opt}(Y, V)$ we need to know γ^0 and, when $q\mathrm{B}(Y, V) \neq 0$, we also need the conditional probabilities $P(X=1|R=1, Y, V)$. However, a feasible double-robust estimator that retains the local efficiency properties of the unfeasible $\widehat{\theta}_{DR}$ can be obtained with the following two-stage algorithm. At the first stage we compute $\widehat{\mu}$ and a preliminary estimator $\widehat{\gamma}$ that solves (2.9) using any $u(Y, V)$. We then calculate the function $\widehat{u}_{opt}(Y, V)$ defined like $u_{opt}(Y, V)$ in (2.10) but with $\widehat{\gamma}$ replacing γ^0 and with the bexpectations in the numerator and denominator computed under $P(X=1|R=1, Y, V; \widehat{\mu})$. At the second stage, we compute $\widehat{\gamma}$ solving (2.9) with u replaced by $\widehat{u}_{opt}$ and finally compute $\widehat{\theta}_{DR}$ using this $\widehat{\gamma}$. This remark raises the interesting point of how much efficiency is gained by the two-stage locally efficient $\widehat{\theta}_{DR}$ compared to the single stage $\widehat{\theta}_{DR}$ computed using $\widehat{\mu}$ and the preliminary $\widehat{\gamma}$. Our simulation study in Section 4 explores this issue using the natural, easy to compute, choice of $u(Y, V) = \partial h(Y, V; \gamma)/\partial\gamma$ for the preliminary estimator $\widehat{\gamma}$. For this choice, the theoretical results of Rotnitzky and Robins, 1997, imply that the single stage and the two-stage estimators are equally asymptotically efficient if $q(Y, V) = 0$ but the two-stage estimator has strictly smaller asymptotic variance if $q(Y, V) \neq 0$. For ease of reference, we call the single stage estimator $\widehat{\theta}_{DR}$ that uses $\widehat{\gamma}$ with $u(Y, V) = \partial h(Y, V; \gamma)/\partial\gamma$ the DR1 estimator and we call the two-stage estimator $\widehat{\theta}_{DR}$ using this $\widehat{\gamma}$ as the preliminary estimator of γ, the DR2 estimator.

Alonzo and Pepe (AP), 2005, derived a locally-efficient double-robust estimator that, except for a minor difference in the computation of $\widehat{\gamma}$, is computed identically to our *DR*2 estimator in the case $q = 0$. Note that when $q = 0$, the function $u_{opt}(Y, V)$ reduces to $\pi\left(Y, V; \gamma^0\right) \times \{\partial h(Y, V; \gamma)/\partial\gamma\}|_{\gamma=\gamma^0}$. The distinction in the computation of $\widehat{\gamma}$ between our procedure and that of AP is that our $\widehat{\gamma}$ solves (2.9) with $u(Y, V)$ replaced by $\pi(Y, V; \widehat{\gamma}) \times \{\partial h(Y, V; \gamma)/\partial\gamma\}|_{\gamma=\widehat{\gamma}}$ while AP's $\widehat{\gamma}$ solves (2.9) but with $u(Y, V)$ replaced by the function $u(Y, V; \gamma) = \pi(Y, V; \gamma) \times \partial h(Y, V; \gamma)/\partial\gamma$, depending on the unknown γ. It is easy to show that both estimators of γ are asymptotically equivalent and efficient. Consequently, when $q = 0$, our DR2 estimator is asymptotically equivalent to AP's estimator of θ.

The following Theorem whose proof is sketched in the Appendix establishes the asymptotic properties of the estimator $\widehat{\theta}_{DR}$ computed using any fixed function $u(Y, V)$. In particular, it

establishes the double-robustness property of $\widehat{\theta}_{DR}$. It also provides a consistent variance estimator when model $\mathscr{A}(q)$, condition (2.7) and either model (2.11) or model (2.8) holds. Further details can be found in Fluss (2006). To state the theorem define $\theta_1 = Se(c)$ and $\theta_2 = Sp(c)$, $\kappa = E(X)$,

$$\begin{aligned}
\Lambda &= -\begin{pmatrix} \kappa & 0 \\ 0 & 1-\kappa \end{pmatrix} \quad \text{and} \quad U(\theta,\gamma,\mu) = \begin{pmatrix} X_{DR}(\gamma,\mu)(T-\theta_1) \\ \{1-X_{DR}(\gamma,\mu)\}(1-T-\theta_2) \end{pmatrix}, \\
\Gamma &= E\left\{\partial U(\theta,\gamma,\mu^*)/\partial\gamma'|_{\gamma=\gamma^*}\right\} E\left\{\partial B(\gamma;u)/\partial\gamma|_{\gamma=\gamma^*}\right\}^{-1}, \\
\Psi &= E\left\{\partial U(\theta,\gamma^*,\mu)/\partial\mu'|_{\mu=\mu^*}\right\} E\left\{\partial H(\mu)/\partial\mu|_{\mu=\mu^*}\right\}^{-1} \quad \text{and} \\
M &\equiv \Lambda^{-1}\{U(\theta,\gamma^*,\mu^*) - \Gamma B(\gamma^*;u) - \Psi H(\mu^*)\}
\end{aligned}$$

where γ^* and μ^* are the probability limits of $\widehat{\gamma}$ and $\widehat{\mu}$. Note that under standard regularity conditions for L-estimators (see for example, van der Vaart (1998)), $\gamma^* = \gamma^0$ when model (2.8) holds and $\mu^* = \mu^0$ when model (2.11) holds.

Theorem 1 *Suppose that model* $\mathscr{A}(q)$ and (2.7) hold. Under standard regularity conditions for L-estimators, if model (2.8) and (2.7) hold or if model (2.11) holds, $\sqrt{n}\left(\widehat{\theta}_{DR} - \theta\right) \to N(0, \Omega)$, where $\Omega = \mathrm{Cov}(M)$. Ω can be consistently estimated with $\widehat{\Omega} = n^{-1}\sum_i \widehat{M}_i\widehat{M}_i'$ *where*

$$\widehat{M} \equiv \widehat{\Lambda}^{-1}\left\{U\left(\widehat{\theta}_{DR},\widehat{\gamma},\widehat{\mu}\right)_i - \widehat{\Gamma}B(\widehat{\gamma};u)_i - \widehat{\Psi}H(\widehat{\mu})_i\right\}$$

$$\begin{aligned}
\widehat{\Gamma} &= \left\{\Sigma_i\, \partial U\left(\widehat{\theta}_{DR},\gamma,\widehat{\mu}\right)_i/\partial\gamma'\Big|_{\gamma=\widehat{\gamma}}\right\}\left\{\Sigma_i\, \partial B(\gamma;u)_i/\partial\gamma|_{\gamma=\widehat{\gamma}}\right\}^{-1}, \\
\widehat{\Psi} &= \left\{\Sigma_i\, \partial U\left(\widehat{\theta}_{DR},\widehat{\gamma},\mu\right)/\partial\mu'\Big|_{\mu=\widehat{\mu}}\right\}\left\{\Sigma_i\, \partial H(\mu)_i/\partial\mu|_{\mu=\widehat{\mu}}\right\}^{-1} \quad \textit{and} \quad \widehat{\Lambda} = -\begin{pmatrix} \widehat{\kappa} & 0 \\ 0 & 1-\widehat{\kappa} \end{pmatrix} \quad \textit{with} \\
\widehat{\kappa} &= \frac{1}{n}\Sigma_i\, X_{DR}(\widehat{\gamma},\widehat{\mu})_i.
\end{aligned}$$

2.1 Illustrative Example

To illustrate our double-robust estimator of the ROC curve we used simulated data following the model used by Rotnitzky *et al.* (2006) in their simulation study. We first generated for each of n=200 subjects a binary disease indicator $X \sim Bernoulli(0.3)$, such that approximately 30% of subjects were diseased. We further simulated a continuous marker Y from the model $Y|X = x \sim N(0.37x, 0.5^2)$, a binary covariate $V = I(X - 0.3 + \varepsilon > 0)$ (and hence conditionally independent of Y) where $\varepsilon \sim N(0, 0.3^2)$, and a response indicator R following the selection for verification model (2.8) with $h(Y, V) = \gamma_0 + \gamma_1 Y + \gamma_2 V$ and $q(Y, V) = \beta$ with $(\gamma_0, \gamma_1, \gamma_2) = (3, -2, -3)$ and $\beta = -1$. The values of $(\gamma_0, \gamma_1, \gamma_2, \beta)$ were chosen so that roughly 70% of the $X's$ would be missing. The covariate V was purposely chosen to have a large correlation (roughly 0.75) with X so that the bias of the estimator based on the verified patients only, be large. The empirical DR1 ROC curves were obtained by estimating $Se(c)$ and $Sp(c)$ as described above for every $c \in \{y_1, ..., y_n\}$ where the y_i's represent the 200 simulated values of Y. Figure 1 illustrates several estimated ROC curves using one simulated sample along with the true and unknown ROC curve (smooth line). For comparison we added the biased estimator (NAIVE) using formula (2.1) based on verified subjects only and the (unfeasible) complete data estimator (COMPLETE) using the entire sample. We also added the estimator assuming missing at random (MAR) using $q = 0$.

Notice the large bias of the NAIVE curve whereas the DR1 is closer to both the complete data estimator and true ROC. DR2 is very similar to DR1 and is not exhibited. The MAR curve is farther from the true ROC than DR1.

3 Comments on DR estimation of the ROC curve

3.1 Monotonicity and Range

The true $Se(c)$ and $1 - Sp(c)$ are non-increasing (decreasing for continuous markers) in c and so the ROC curve is a monotonic non-decreasing (increasing for continuous markers) function of $1 - Sp(c)$. However, as noticeable in our example (Figure ??), the DR estimate of the ROC curve is not necessarily non-decreasing. Further, the ROC curve can attain values out of the possible range of [0, 1]. Both are due to the possibility of $X_{DR}(\widehat{\gamma},\widehat{\mu})_i$ being outside of the interval [0, 1].

Non-monotonicity can be corrected by applying an isotonic regression on the estimated sensitivity and specificity. For the sensitivity, the isotonic regression estimator is defined as follows. Let $\widehat{\mathbf{Se}}$ be the $n \times 1$ vector with j^{th} coordinate $\widehat{\mathbf{Se}}_j$ equal to $\widehat{Se}(c_j)$, the double-robust estimator of $Se(cj)$. For any non-increasing function $f(\cdot)$ on the reals, define the $n \times 1$ vector $\mathbf{f}$ whose j^{th} is equal to $f(c_j)$. Define

$$\widehat{\mathbf{f}} = \arg \min_{\mathbf{f}: f \text{ is non-increasing}} \left\| \widehat{\mathbf{Se}} - \mathbf{f} \right\|^2$$

where for any $n \times 1$ vector u, $\|\mathbf{u}\|^2 = \sum_{i=1}^{n} u_i^2$. The isotonic regression estimator of $Se(c_j)$ is defined as $\widehat{Se}(c_j)_{iso} = \widehat{\mathbf{f}}_j$ (Barlow *et al.*, 1978). As indicated by these authors, the restricted minimization needed to compute $\widehat{\mathbf{f}}$, can be carried out using the pooled-adjacent-violators (PAV) algorithm. Following the theory in Barlow *et al.* (1978, Theorem, 2,2) it can be shown that convergence in probability of $\widehat{\mathbf{Se}}_j$ to $\mathbf{Se}_j$ for all $j = 1, ..., n$, implies the consistency of $\widehat{\mathbf{Se}}_{iso,j}$ as an estimator $\mathbf{Se}_j$ for each j. The isotonic regression is commonly recommended for adjusting non-monotonic estimators of monotonic functions. For example, see Jewell and Van der Laan (2004) in the context of estimating the survival function estimation with censored data and Alonzo and Pepe (2005) for estimation of the ROC curve under verification bias assuming a MAR process.

After applying the PAV algorithm to both the estimated sensitivity and specificity, these resulting estimates can still be outside the range [0, 1]. We correct for this by adjusting the DR estimates greater than 1 (or less than 0) to be 1 (or 0). We refer to these adjusted estimates as DR-PAV. In Figure 2 we present the DR-PAV estimated ROC curve of the illustrative example (Section 2.1). The general form of the ROC curve remains the same but the curve is much smoother now. Formulae for the asymptotic variance of the DR-PAV estimate is not presently available. However, we speculate that the non-parametric bootstrap estimator of variance is a consistent estimator of the asymptotic variance of $\widehat{Se}(c)_{iso}$. The simulation results presented in Section 4 support our speculation. Further theoretical investigations are needed to confirm this speculation.

3.2 Confidence Intervals for the ROC Curve

After computing the estimator $\widehat{\Omega}$ of the asymptotic variance-covariance matrix of $\widehat{\theta}_{DR}$ (not applying the isotonic regression) we can obtain marginal $1 - \alpha$ confidence intervals (CIs) for $Se(c)$ and $Sp(c)$ for a fixed value of c. A joint confidence region (CR) for bthe pair ($Se(c)$,

$Sp(c)$) can be constructed either by applying the Bonferroni correction (BON) or by the elliptical CR (MVN) based on the asymptotic bivariate normality of $\widehat{\theta}_{DR}$. Applying the logit(·) transformation on $\widehat{\theta}_{DR}$ usually improves the normality of these estimators and as a result may improve coverage in both marginal and joint CIs. The asymptotic variance-covariance matrix of $\text{logit}\left(\widehat{\theta}_{DR}\right)$ can be obtained using the delta method. Fluss (2006) further discusses CIs for the DR estimated ROC curves and provides some simulation results which indicate that for large sample sizes (n=2000) the coverage reaches the nominal level and the logit transformation makes no real difference. For small sample sizes (n=200) the CIs usually exhibit undercoverage but the logit transformation improves coverage. The Bonferroni and the bivariate normal methods had similar performances. An example of these regions is given below (Section 5).

Alternatively a confidence band (CB) for the entire curve as a whole can be obtained. Campbell (1994) suggests using bootstrap methods to estimate the distribution of the maximum distance between the true and the estimated ROC curves as a basis for constructing the CB. In Fluss (2006) a detailed bootstrap procedure is given.

4 Simulation study

We conducted a simulation study to examine the finite sample behavior of $\widehat{\theta}_{DR}$. We examined the behavior of the DR1 and DR2 estimators under correct and incorrect working models. We used the model described in Section (2.1), generating 1000 replicates with sample sizes $n = 200$ and $n = 2000$, under two scenarios: (a) $\beta = -1$ and (b) $\beta = 0$. For each simulated data set we estimated ($Se(c)$, $Sp(c)$) for $c = 0.4208$ (corresponding to $Se(c) = 0.4595$ and $Sp(c) = 0.8$). For the DR2 estimator, we solved (2.9) with $\widehat{u}_{opt}(Y,V) = (1,Y,V)' \times \widehat{E}\{\exp(\beta X) \mid R=1,Y,V\} \times \widehat{E}\{\exp(\beta X)\,\pi(Y,V;\widehat{\gamma}^{-}) \mid R=1,Y,V\}^{-1}$ (see (2.10)) where $\widehat{E}$ denotes the conditional expectation under the working disease model substituting $\widehat{\mu}$ for μ and $\widehat{\gamma}^{-}$ for γ^0 where $\widehat{\gamma}^{-}$ is the estimator of γ used to computed the DR1 estimator (i.e. $\widehat{\gamma}^{-}$ solving (2.9) with $u(Y, V) = \partial h(Y, V; \gamma)/\partial$). We also examined DR1 and DR2 after applying the PAV procedure (Section 3.1). For comparison we considered also the Monte-Carlo behavior of the inconsistent estimator (NAIVE) using only verified subjects in formula (2.1), the complete data estimator (COMPLETE) assuming all subjects in the sample are verified. In scenario (a) we also examined the impact of incorrectly assuming that $\beta = 0$ and thus computed the estimator DR2 using $\beta = 0$ (MAR). We also computed the estimator of Alonzo-Pepe, 2005, but we do not report it here because its Monte Carlo performance was nearly identical to that of the MAR estimator, confirming the theoretical results that establish that the estimators are asymptotically equivalent.

In our simulations we computed $\widehat{\theta}_{DR}$ under the following four scenarios for the working models:

- **i.** Both selection for verification and disease models are correctly specified. The model for selection for verification is: $logit(\Pr(R = 1 \mid Y, V, X)) = \gamma_0 + \gamma_1 Y + \gamma_2 V + \beta$ and the disease model is: $logit(\Pr(X = 1 \mid Y, V, R = 1)) = \log\{o(Y, V)\} + \mu_0 + \mu_1 Y + \mu_2 V$, where $o(Y,V) = \frac{[1+\exp\{\gamma_0+\gamma_1 Y+\gamma_2 V+\beta\}]}{1+\exp\{\gamma_0+\gamma_1 Y+\gamma_2 V\}}$. For calculating $o(Y, V)$ we used the correct values of γ.
- **ii.** Only the selection for verification model is correctly specified. The disease model was incorrectly specified in that we forced $\mu_1 = 0$ and $o(Y, V) = 1$.

iii. Only the disease model is correctly specified. The model for selection for verification was incorrectly specified in that we set $\gamma_1 = 0$.

iv. Both working models are incorrectly specified as in (*ii*) and (*iii*).

It can be shown that DR2 and DR1 are algebraically identical when they are computed under the working models considered in (*iv*) . They are also identical under the setting (*iii*) if the estimator $\widehat{\theta}_{DR}$ assumes $\beta = 0$ (see proof in Fluss (2006)). The performances of the estimators of $\theta(c)$ were assessed by means of Monte Carlo bias and standard error (MCSE). In addition, we examined the agreement between the Monte Carlo mean of the estimated standard errors (computed with the formulae given in Theorem 1) and the MC standard error of the DR estimators. As formulae for estimates of the standard error of the DR-PAV estimators are not presently available, for these estimators we report results on the bootstrap estimator of standard error. Due to time considerations the bootstrap was only computed for DR2 and $n = 200$. The symbol '–' stands for unavailable data.

In our simulations, applying the PAV algorithm generally had little effect on both the bias and standard error so here we will restrict our discussion to the simulation results for the non-PAV adjusted estimators. As predicted by theory, the simulation results show that the DR estimators greatly corrected the noticeable bias of the NAIVE estimator (especially when n=2000) in any of the first three scenarios in which either the disease or the verification processes were correctly modeled. Even when both working models were incorrectly specified (scenario (*iv*)) the DR method exhibited substantially smaller bias than that of the NAIVE estimator. When the true β was −1 and $n = 2000$, the MAR/Alonzo-Pepe estimator had bias that, even though small (roughly 7% relative bias for *Se* and 0.5% for *Sp* in scenarios (*i*) and (*ii*)), it was orders of magnitude greater than that of the DR estimator that used the true value of β. When $n = 200$, and under the same scenarios (*i*) and (*ii*) , the biases of the DR estimators were still smaller than those of the MAR estimators but of comparable sizes. In scenario (*iii*) the bias reduction of the DR estimator compared to the MAR estimator still existed but was less pronounced. Finally, in scenario (*iv*) in which, according to the theory none of the estimators is consistent, we see that for estimation of *Se*, the magnitude of the bias is still smaller for the DR compared to that of the MAR estimator but this order reverses for estimation of *Sp*. The reason why under scenarios (*i*) – (*iii*) the MAR estimator is biased downwards is as follows. Under our setting, for subjects with a given value of Y and V, the odds of being verified are 2.7 times greater for those diseased than for those non-diseased. Thus, within each level of Y and V the fraction of non-verified subjects will be larger in the non-diseased group than in the diseased group. However, the procedure that incorrectly assumes MAR will implicitly impute, within levels of V and Y, half of the non-verified to the diseased group and the other half to the non-diseased group. But this will spuriously inflate the left tail of the distribution of Y for disease subjects and deflate the distribution of Y for the non-diseased subjects because most of the imputation will occur for low values of Y since the chances of being verified increase with Y. The effect of this spurious distortion of the tail distribution is that both $P(Y \le c|X = 0)$ and $P(Y > c|X = 1)$ will be underestimated. Furthermore, because the diseased subjects are only 30% of the entire sample, the distortion caused on the distribution of Y by incorrectly assigning non-diseased subjects to the non-disease group will be greater on the diseased group than on the non-diseased group. Thus, we would expect (as confirmed in our simulations) the bias in the estimation of $Se = P(Y > c|X = 1)$ to be greater than the bias in the estimation of $Sp = P(Y \le c|X = 0)$.

The MCSE was similar to the Monte Carlo mean of the estimated standard errors, especially for n=2000. For n=200, the estimator of standard error underestimated a bit. As predicted by the theory when $\beta = -1$ and both working models are correct, i.e. under scenario (*i*) , the Monte-Carlo variance of the DR2 estimator was less than that of the DR1 estimator (the

variance reduction being, nevertheless, generally small). When both working models are correct and $\beta = 0$, the theoretical results establish that both DR1 and DR2 are asymptotically equivalent. This was confirmed in our simulations for $n = 2000$ since the MCSE of DR1 and DR2 were roughly the same. Yet, for estimation of Se when $n = 200$, DR2 still exhibited some variance reduction over that of DR1 (roughly 5% reduction).

The comparison of the MCSE of the COMPLETE and DR2 estimators raises another interesting point. The COMPLETE estimator was calculated using all n simulated data points (i.e. as though no $X's$ were missing). Yet, in our analysis, roughly 70% of the subjects had missing X. An estimator based just on the data points with X observed, would then be expected to have MC variance roughly equal to 1/0.3=3.33 times the MC variance of the COMPLETE estimator. Yet, the MC variance of the DR2 estimator was far smaller than that. For example, in scenario 1 and $\beta = -1$, the $(MCSE(DR2)/MCSE(COMPLETE))^2 = 1.57$. This result confirms the theoretical efficiency properties of the DR2 procedure. Specifically, the estimator that uses the DR2 estimator with both working models correctly specified, exploits *all* the information available in the observed data about the parameters Sp and Se. Thus, it not only exploits the information in the units with observed $X's$ but also in the units with missing $X's$.

When examining the Monte Carlo sampling distributions of the DR estimators we find them reasonably bell-shaped but with a few extreme values. Fluss (2006) considers additional choices of β and c, and different disease and selection models. Due to space constraints, these simulations are not presented here. They nevertheless were qualitatively similar to those presented here.

5 EBCT data

We apply the double-robust procedures of Section 2 to the data example described in the introduction with the goal of estimating the ROC curve of the (log-transformed) EBCT marker. To that end, we define $Y = \ln(1 + EBCT)$, X the indicator of coronary disease as determined by the SPEC test (regarded here as the gold standard) and $V = (V_1, V_2)$ a bivariate vector comprised by age (V_1) and the indicator of aspirin use (V_2). We consider the working verification and disease models

$$\begin{aligned} h(Y, V; \gamma) &= \gamma_0 + \gamma_1 I(Y=0) + \gamma_2 I(Y \neq 0)\, Y + \gamma_3 I(Y \neq 0)\, V_2 . \\ m(Y, V; \mu) &= \mu_0 + \mu_1 Y + \mu_2 V_1 + \mu_3 V_2 . \end{aligned}$$

Similar working models were considered in Rotnitzky *et al.* (2006) who give the rationale for these choices. In the analysis reported below we use $q(Y, V) = \beta$, repeating the analysis for values of β regarded as known ranging from −2 to 0. This range was chosen so as to include settings with very severe residual selection bias in favor of diseased subjects (the choice $\beta = -2$ indicates that the odds of selection is $\exp(2) \approx 7$ times larger for diseased than healthy patients with the same values of EBCT, age and aspirin use), and ignorable verification ($\beta = 0$). The estimated ROC curves using DR2 (applying the PAV algorithm) are presented in Figure 3. The results without PAV are very similar, as found in the simulation study for large sample sizes, and are not presented. The DR1 estimate provides similar results and is not given. The estimated ROC curve based on the verified only subjects is also presented labelled as NAIVE.

The naive estimate gives a significantly different ROC curve than the DR and with a smaller area under the curve. We can see there is an evident effect of the value of β on the ROC curve, but all the DR estimated curves show the marker has better diagnostic power than is shown by the naive estimator. We constructed 95% confidence regions for several points on

the ROC curve. Three points corresponding to $\widehat{S}p_{DR}(c)=0.7, 0.8$ and 0.9 (using $\beta=-1$) are presented in Figure 4. Since Fluss (2006) found the logit transformation did not make much of a difference for large sample size we used the simple Bonferroni (BON) and Multivariate Normal (MVN) distribution based CR. The resulting CRs are shown in Figure 4. The MVN CRs are very similar to the Bonferroni ones. The width of the CRs in the sensitivity axes is very wide indicating on a large amount of uncertainty in the sensitivity estimation. This is due to the small number of verified disease cases.

The estimated ROC curves adjusted by our procedure for verification bias show a higher discriminatory power than the ROC curve based on verified only subjects. This holds true for a wide range of residual selection bias values. An advantage of our method is that it permits the examination of the consistency of qualitative conclusions on the adjusted ROC curves for selection bias ranging from negligible to considerable. We note from Figure 3 that for high specificity values the corresponding sensitivities are only moderate in size. Adding to this consideration the large range of values for sensitivity in the CRs of Figure 4 it is difficult to make a definite conclusion as to the effectiveness of the EBCT marker. An anonymous referee pointed out that SPECT perfusion imagining is not really a gold standard for coronary artery disease. In fact, angiography is the accepted gold standard. The medical literature supports the findings that the specificity and sensitivity of SPECT for angiographically confirmed CAD are less than 0.9. This may explain, in part, the somewhat disappointing results for EBCT.

Acknowledgments

This research was supported by a grant from the United States - Israel Binational Science Foundation (BSF), Jerusalem, Israel. In addition, Andrea Rotnitzky was partially funded by grant R01-GM48704 from the National Institutes of Health, U.S.

The authors gratefully acknowledge the comments of the editor, associate editor and two reviewers, which led to improvements in the paper.

Appendix: Sketch of the Proof of Theorem 1

Regardless of whether or not models (2.8) and (2.11) are correct, the solutions $\widehat{\mu}$ and $\widehat{\gamma}$ to $\sum_i H(\mu)_i=0$ and $\sum_i B(\gamma;u)_i=0$ converge in probability to μ^* and γ^* solving $E\{H(\mu^*)\}=0$ and $E\{B(\gamma^*;u)\}=0$. Standard Taylor expansions arguments give

$$\begin{aligned}\sqrt{n}(\widehat{\mu}-\mu^*) &= -E\left\{\partial H(\mu)/\partial\mu|_{\mu=\mu^*}\right\}^{-1}\sqrt{n}\mathbb{P}_n\{H(\mu^*)\}+o_p(1)\\ \sqrt{n}(\widehat{\gamma}-\gamma^*) &= -E\left\{\partial B_i(\gamma;u)/\partial\gamma|_{\gamma=\gamma^*}\right\}^{-1}\sqrt{n}\mathbb{P}_n\{B(\gamma^*;u)\}+o_p(1)\end{aligned}$$

where for any random variables $W_1,\ldots,W_n$, $\mathbb{P}_n(W)$ stands for $n^{-1}\sum_{i=1}^n W_i$. Another Taylor expansion gives

$$\begin{aligned}&\sqrt{n}\mathbb{P}_n\left\{U\left(\widehat{\theta}_{DR},\widehat{\gamma},\widehat{\mu}\right)\right\}=0\\ =\;&\sqrt{n}\mathbb{P}_n\{U(\theta,\gamma^*,\mu^*)\}+\left[\frac{\partial}{\partial\theta}\mathbb{P}_n\{U(\theta,\gamma^*,\mu^*)\}|_{\theta=\bar{\theta}}\right]\sqrt{n}\left(\widehat{\theta}_{DR}-\theta\right)\\ &+\left[\frac{\partial}{\partial\gamma}\mathbb{P}_n\{U(\theta,\gamma,\mu^*)\}|_{\gamma=\bar{\gamma}}\right]\sqrt{n}(\widehat{\gamma}-\gamma^*)+\left[\frac{\partial}{\partial\mu}\mathbb{P}_n\{U(\theta,\gamma^*,\mu)\}|_{\mu=\bar{\mu}}\right]\sqrt{n}(\widehat{\mu}-\mu^*)\end{aligned}\quad(A.1)$$

for some $\bar{\theta}$, $\bar{\gamma}$ and $\bar{\mu}$ satisfying $\|\bar{\theta}-\theta\|\le\|\widehat{\theta}-\theta\|=o_p(1)$, $\|\bar{\gamma}-\gamma^*\|\le\|\widehat{\gamma}-\gamma^*\|=o_p(1)$ and $\|\bar{\mu}-\mu^*\|\le\|\widehat{\mu}-\mu^*\|=o_p(1)$. Now,

$$\frac{\partial}{\partial\theta}\mathbb{P}_n\{U(\theta,\gamma^*,\mu^*)\}\big|_{\theta=\bar{\theta}} = -\begin{pmatrix} \mathbb{P}_n\{X_{DR}(\gamma^*,\mu^*)\} & 0 \\ 0 & \mathbb{P}_n\{1-X_{DR}(\gamma^*,\mu^*)\} \end{pmatrix} = \Lambda \quad \text{(A.2)}$$

where the last identity follows from the Law of Large Numbers and the fact that $E\{X_{DR}(\gamma^*,\mu^*)\,|\,Y,V\} = E\{X|Y,V\}$ when one of models (2.8) or (2.11) is correct, but not necessarily both. Furthermore, under regularity conditions, a Uniform Law of Large Numbers gives $\partial\mathbb{P}_n\{U(\theta,\gamma,\mu^*)\}/\partial\gamma|_{\gamma=\bar{\gamma}} = E\left\{\partial U(\theta,\gamma,\mu^*)/\partial\gamma|_{\gamma=\gamma^*}\right\}$ and $\partial\mathbb{P}_n\{U(\theta,\gamma^*,\mu)\}/\partial\mu|_{\mu=\bar{\mu}} = E\left\{\partial U(\theta,\gamma^*,\mu)/\partial\mu|_{\mu=\mu^*}\right\}$. Consequently, replacing the expansions (A.1) and the latter expressions for the derivatives in (A.2) and solving for $\sqrt{n}\left(\widehat{\theta}_{DR}-\theta\right)$ we obtain

$$\sqrt{n}\left(\widehat{\theta}_{DR}-\theta\right) = \sqrt{n}\mathbb{P}_n(M) + o_p(1)$$

and the result follows after invoking the Central Limit Theorem. The condition (2.7) is required to ensure that $B(\gamma^*, u)$ and $X_{DR}(\gamma^*,\mu^*)$ have finite variance when model (2.8) holds but model (2.11) is incorrect.

References

1. ALONZO TA, PEPE MS. Assesing accuracy of a continuous screening test in the presence of verification bias. Applied Statistics. 2005; 54(1):173–190.
2. BAKER SG. Evaluating multiple diagnostic tests with partial verification. Biometrics. 1995; 51:330–337. [PubMed: 7539300]
3. BARLOW, RE.; BARTHOLOMEW, DJ.; BREMNER, JM.; BRUNK, HD. Statistical Inference under Order Restrictions. Wiley; New York: 1978.
4. BEGG C, GREENES R. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics. 1983; 39:207–215. [PubMed: 6871349]
5. BRAUN J, OLDENDORF M, MOSHAGE W. Electron beam computed tomography in the evaluation of cardiac classifications in chronic dialysis patients. American Journal of Kidney Diseases. 1996; 27:394–401. [PubMed: 8604709]
6. CAMPBELL G. Advances in statistical methodology for the evaluation of diagnostic and laboratory tests. Statistics in Medicine. 1994; 13:499–508. [PubMed: 8023031]
7. FLUSS, R. Estimation of the ROC Curve and its Associated Indices Under Verification Bias. University of Haifa; 2006. unpublished doctoral dissertation
8. GRAY R, BEGG C, GREENES R. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. Medical Decision Making. 1984; 4:151–164. [PubMed: 6472063]
9. JEWELL, NP.; VAN DER LAAN, MJ. Advances in Survival Analysis.. In: BALAKRISHNAN, N.; RAO, CR., editors. Handbook of Statistics. Vol. 23. Elsevier North Holland; 2004. p. 625-643.
10. KOSINSKI AS, BARNHART HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. Biometrics. 2003; 59:163–171. [PubMed: 12762453]
11. LITTLE, RJ.; RUBIN, DB. Statistical analysis with missing data. Wiley; New York: 1987.
12. ROBINS, JM.; ROTNITZKY, A.; SCHARFSTEIN, DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models.. In: HALLORAN, ME.; BERRY, D., editors. Statistical Models for Epidemiology, the Environment, and Clinical Trials. Springer-Verlag; New York: 2000. p. 1-95.
13. RODENBERG C, Zhou XH. ROC curve estimation when covariates affect the verification process. Biometrics. 2000; 56:131–136.

14. ROTNITZKY A, ROBINS J. Analysis of semi-parametric regression models with non-ignorable non-response. Statistics in Medicine. 1997; 16:81–102. [PubMed: 9004385]

15. 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.

16. RUBIN D. Inference and missing data. Biometrika. 1976; 72:581–592.

17. SCHARFSTEIN DO, ROTNITZKY A, ROBINS JM. Adjusting for nonignorable drop-out using semiparametric nonresponse models (with discussion). Journal of the American Statistical Association. 1999; 94:1096–1120.

18. TOLEDANO A, GATSONIS C. Ordinal regression methodology for ROC curves derived from correlated data. Statistics in Medicine. 1996; 15:1807–1826. [PubMed: 8870162]

19. VAN DER VAART, AW. Asymptotic statistics. Cambridge University Press; 1998.

20. ZHOU XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. Communications in Statistics - Theory and Methods. 1993; 54:124–135.

21. ZHOU XH. Effect of verification bias on positive and negative predictive values. Statistics in Medicine. 1994; 13:1737–1745. [PubMed: 7997707]

22. ZHOU XH. A nonparametric maximum likelihood estimator for the receiver operating characteristic curve area in the presence of verification bias. Biometrics. 1996; 52:299–305. [PubMed: 8934599]

23. ZHOU XH. Correcting for verification bias in studies of a diagnostic test's accuracy. Statistical Methods in Medical Research. 1998a; 7:337–353. [PubMed: 9871951]

24. ZHOU XH. Comparing Accuracies of Two Screening Tests in the Presence of Verification Bias. Journal of the Royal Statistical Society, Series C. 1998b; 47:135–147.

25. ZHOU XH. Comparing the correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias. Biometrics. 1998c; 54:453–470. [PubMed: 9629639]

26. ZHOU XH, RODENBERG C. Estimating the ROC curve in the presence of nonignorable verification bias. Communications in Statistics - Theory and Methods. 1998; 27:635–57.

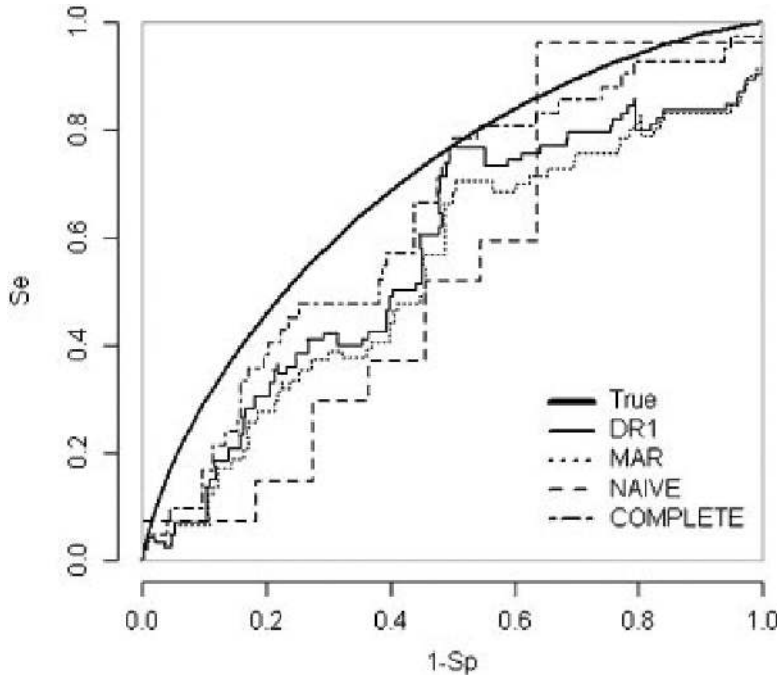

Figure 1.
Example of DR estimated ROC curves.

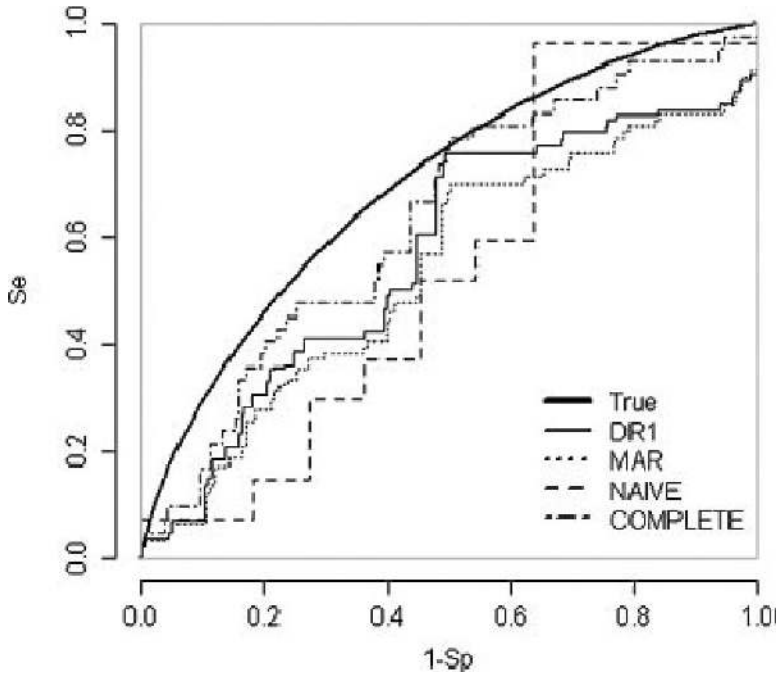

Figure 2.
Estimated ROC curves of example after applying PAV.

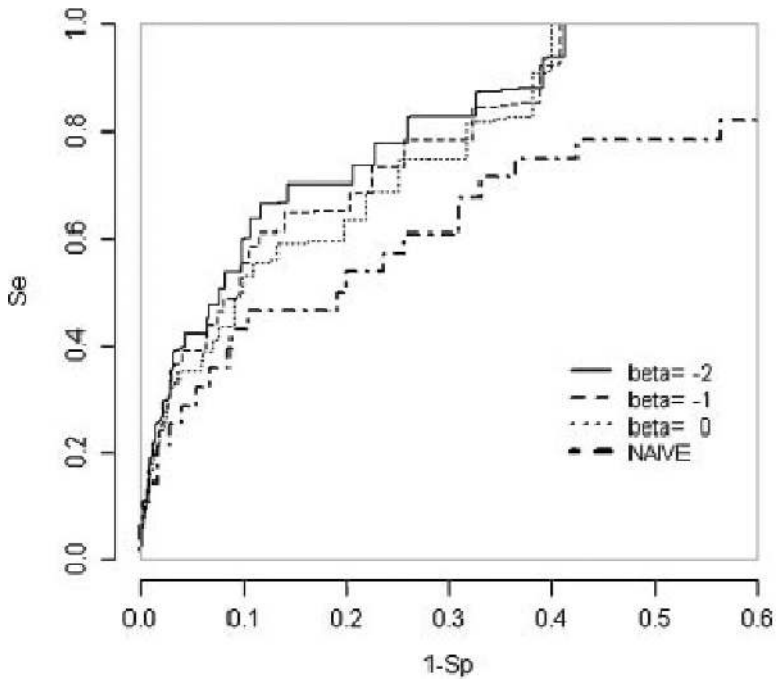

Figure 3.
DR estimated ROC curves of the EBCT example.

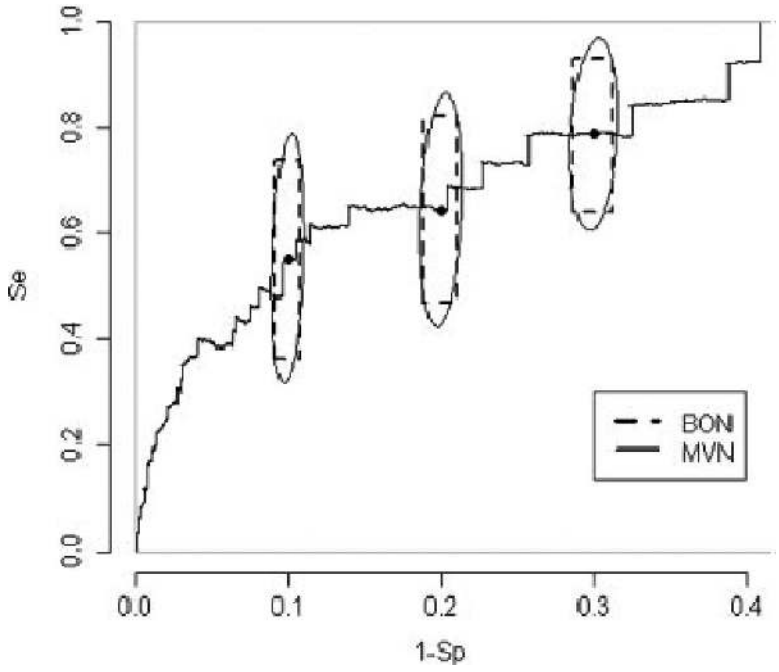

Figure 4.
CR for 3 points on the EBCT ROC curve

Table 1

Simulation Study Results: (Se, Sp) = (0.4595, 0.8), $\beta = -1$. All values have been multiplied by 100.

Working Models	Method	n=200 Bias	n=200 Estimated SE	n=200 MCSE	n=2000 Bias	n=2000 Estimated SE	n=2000 MCSE
Sensitivity							
i	DR1	−1.00	7.52	8.86	0.14	2.51	2.58
i	DR1-PAV	−0.98	-	8.84	0.14	-	2.58
i	DR2	−0.30	7.72	8.09	0.17	2.49	2.58
i	DR2-PAV	−0.28	7.90	8.06	0.17	-	2.57
i	MAR	−3.62	-	7.82	−3.52	-	2.32
ii	DR1	−0.17	7.93	9.88	0.30	2.62	2.68
ii	DR1-PAV	−0.15	-	9.85	0.30	-	2.68
ii	DR2	−0.53	7.83	8.35	0.15	2.58	2.67
ii	DR2-PAV	−0.51	7.98	8.33	0.15	-	2.67
ii	MAR	−3.12	-	8.68	−3.54	-	2.32
iii	DR1	−0.05	8.00	8.16	0.20	2.56	2.63
iii	DR1-PAV	−0.02	-	8.12	0.20	-	2.63
iii	DR2	−0.05	7.97	8.12	0.20	2.55	2.62
iii	DR2-PAV	−0.02	7.98	8.08	0.19	-	2.62
iii	MAR	−5.4	-	7.47	−5.77	-	2.42
iv	DR1/2	−1.21	7.18	7.27	−0.79	2.28	2.36
iv	DR1/2-PAV	−1.18	7.16	7.24	−0.79	-	2.35
iv	MAR	−4.62	-	6.69	−4.24	-	2.16
-	NAIVE	11.59	-	8.20	12.21	-	2.47
-	COMPLETE	−0.48	-	6.44	0.03	-	2.07
Specificity							
i	DR1	−0.42	3.63	3.92	0.01	1.14	1.11
i	DR1-PAV	−0.42	-	3.92	0.01	-	1.11
i	DR2	−0.15	3.63	3.79	0.01	1.14	1.10
i	DR2-PAV	−0.15	3.81	3.78	0.01	-	1.10
i	MAR	−0.45	-	3.89	−0.25	-	1.19

Sensitivity

Working Models	Method	n=200			n=2000		
		Bias	Estimated SE	MCSE	Bias	Estimated SE	MCSE
i	DR1	−0.36	3.61	3.90	0.02	1.14	1.11
ii	DR1-PAV	−0.36	-	3.90	0.02	-	1.11
ii	DR2	−0.19	3.63	3.77	0.01	1.14	1.12
i	DR2-PAV	−0.19	3.77	3.77	0.01	-	1.12
ii	MAR	−0.40	-	3.86	−0.26	-	1.19
iii	DR1	−0.06	3.73	3.86	0.03	1.17	1.14
iii	DR1-PAV	−0.06	-	3.86	0.03	-	1.14
iii	DR2	−0.06	3.71	3.85	0.03	1.17	1.14
iii	DR2-PAV	−0.06	3.83	3.85	0.03	-	1.14
iii	MAR	−1.11	-	4.31	−1.23	-	1.31
iv	DR1/2	0.27	3.95	4.08	0.36	1.25	1.21
iv	DR1/2-PAV	0.27	3.98	4.08	0.36	-	1.21
iv	MAR	−0.13	-	4.38	−0.01	-	1.34
-	NAIVE	−24.58	-	12.72	−25.68	-	4.02
-	COMPLETE	−0.20	-	3.59	−0.01	-	1.06

Table 2

Simulation Study Results: (Se, Sp) = (0.4595, 0.8), $\beta = 0$. All values have been multiplied by 100.

Working Models	Method	n=200 Bias	n=200 Estimated SE	n=200 MCSE	n=2000 Bias	n=2000 Estimated SE	n=2000 MCSE
Sensitivity							
i	DR1	0.35	7.30	7.84	0.12	2.38	2.53
i	DR1-PAV	0.34	-	7.82	0.12	-	2.53
i	DR2	0.17	7.28	7.47	0.11	2.37	2.50
i	DR2-PAV	0.16	7.51	7.45	0.11	-	2.50
ii	DR1	1.38	7.94	8.91	0.35	2.48	2.68
ii	DR1-PAV	1.37	-	8.89	0.35	-	2.68
ii	DR2	0	7.49	7.65	0.10	2.42	2.56
ii	DR2-PAV	–0.02	7.83	7.63	0.10	-	2.56
iii	DR1	0.15	7.41	7.76	0.11	2.45	2.54
iii	DR1-PAV	0.13	7.52	7.72	0.11	-	2.54
iv	DR1	–1.23	6.59	6.91	–1.10	2.17	2.28
iv	DR1-PAV	–1.25	6.65	6.86	–1.10	-	2.28
-	NAIVE	11.99	-	7.94	12.11	-	2.52
-	COMPLETE	–0.08	-	6.54	0.07	-	2.10
Specificity							
i	DR1	0.01	3.49	3.36	–0.02	1.12	1.17
i	DR1-PAV	0.01	-	3.36	–0.02	-	1.17
i	DR2	0.02	3.52	3.40	–0.03	1.12	1.17
i	DR2-PAV	0.01	3.56	3.40	–0.03	-	1.17
ii	DR1	0.10	3.49	3.35	0	1.13	1.17
ii	DR1-PAV	0.10	-	3.35	0	-	1.17
ii	DR2	–0.01	3.54	3.40	–0.03	1.13	1.17
ii	DR2-PAV	–0.01	3.56	3.40	–0.03	-	1.17
iii	DR1	0.06	3.61	3.61	–0.02	1.17	1.21
iii	DR1-PAV	0.06	3.61	3.61	–0.02	-	1.21
iv	DR1	0.33	3.83	3.86	0.27	1.24	1.29

Sensitivity								
			n=200			n=2000		
Working Models	Method	Bias	Estimated SE	MCSE	Bias	Estimated SE	MCSE
iv	DR1-PAV	0.32	3.78	3.83	0.27	-	1.29
-	NAIVE	−19.33	-	11.34	−19.98	-	3.48
-	COMPLETE	0	-	3.25	−0.02	-	1.09