Published in final edited form as:

Comput Stat Data Anal. 2009 January 15; 53(3): 707–717. doi:10.1016/j.csda.2008.06.021.

Estimation of the disease-specific diagnostic marker distribution under verification bias

John H. Page¹ and Andrea Rotnitzky²

¹ Department of Epidemiology, Harvard School of Public Health, Boston, USA

² Department of Economics, Di Tella University, Buenos Aires, Argentina, and Department of Biostatistics, Harvard School of Public Health, Boston, USA

Abstract

We consider the estimation of the parameters indexing a parametric model for the conditional distribution of a diagnostic marker given covariates and disease status. Such models are useful for the evaluation of whether and to what extent a marker's ability to accurately detect or discard disease depends on patient characteristics. A frequent problem that complicates the estimation of the model parameters is that estimation must be conducted from observational studies. Often, in such studies not all patients undergo the gold standard assessment of disease. Furthermore, the decision as to whether a patient undergoes verification is not controlled by study design. In such scenarios, maximum likelihood estimators based on subjects with observed disease status are generally biased. In this paper, we propose estimators for the model parameters that adjust for selection to verification that may depend on measured patient characteristics and additionally adjust for an assumed degree of residual association. Such estimators may be used as part of a sensitivity analysis for plausible degrees of residual association. We describe a doubly robust estimator that has the attractive feature of being consistent if either a model for the probability of selection to verification or a model for the probability of disease among the verified subjects (but not necessarily both) is correct.

Keywords

Missing at Random; Nonignorable; Missing Covariate; Sensitivity Analysis; Semiparametric; Diagnosis

1 Introduction

In evaluating the usefulness of a diagnostic marker an important question is whether and to what extent a marker's diagnostic ability depends on the patient's characteristics. To that extent, models for the distribution of the marker conditional on covariates and disease status play an important role in that they help elucidate this dependence. For instance, these models can be used to compute covariate specific measures of diagnostic accuracy such as, specificity, sensitivity, the receiver operating characteristic curve and the area under the curve (AUC) or the partial area, the Youden index and the likelihood ratio.

© 2008 Elsevier B.V. All rights reserved.

Publisher's Disclaimer: This is a PDF file of an unedited manuscript that has been accepted for publication. As a service to our customers we are providing this early version of the manuscript. The manuscript will undergo copyediting, typesetting, and review of the resulting proof before it is published in its final citable form. Please note that during the production process errors may be discovered which could affect the content, and all legal disclaimers that apply to the journal pertain.

Estimation of diagnostic marker accuracy often relies on observational databases of medical records. Because disease verification is often expensive and/or invasive, doctors send their patients for the definite assessment of disease only if, based on the marker results and patient's risk factors, they judge it necessary to do so. Consequently, verified subjects often differ from non-verified subjects in prognostic factors for disease. As such, estimators of diagnostic accuracy measures based on verified subjects only are generally biased.

A number of articles have proposed methods to correct the estimation of diagnostic test accuracy measures under verification bias, most of which are reviewed in the book of Zhou, Obuchowski, and McClish, 2002 and in Zhou, 1998a. The articles of Rodenberg and Zhou, 2000, Gray, Begg and Greenes, 1984, Zhou, 1998b, and Zhou and Riggs, 2000, Toledano and Gatsonis 1996, 1999, Zhou, 1996 considered only discrete covariates and assumed ignorable verification, i.e. that the probability of verification is conditionally independent of true disease status given the test result and covariates. However, ignorable verification is often an unrealistic assumption because the doctor's decision to send a patient to verification is typically based on his/her detailed information on the patient's health status, which cannot be accurately summarized by measured covariates, much less by a discrete covariate. Acknowledging this, Baker, 1995, Kosinski and Barnhart, 2003, and Zhou, 1993, 1994, Zhou and Rodenberg, 1998 and Zhou and Castelluccio, 2003 considered estimation under non-ignorable verification models. An important limitation of these papers is that they assumed that both the covariates and marker are discrete variables. Recently, Rotnitzky, Faraggi, and Schisterman, 2005, described estimators of the AUC under ignorable or non-ignorable verification bias, which contemplate the possibility that the marker is continuous and/or that the covariates are continuous and high dimensional.

In this paper we derive estimators of the parameters indexing a parametric model for the conditional distribution of a continuous or ordinal marker given disease status and, a possibly high dimensional vector of, covariates when the disease status is observed only on a subset of the sample. Our estimators adjust for selection to verification that may depend on measured patient characteristics and marker values, and additionally adjust for an assumed degree of residual association between selection to verification and true disease status due to unmeasured patient variables.

Our research was motivated by the investigation of the diagnostic properties of a cardiology device called electron beam computed tomography (EBCT). EBCT is a device routinely used to measure the degree of calcification in the arteries. It is unknown, however, if EBCT can also help screen for myocardial ischemia (reduced blood flow due to obstruction in the vessels) and if so, to what extent its screening ability depends on the patient's risk factors. These are important questions because the other available noninvasive instrument for screening myocardial ischemia is the Dual Isotope Myocardial Perfusion Single Photon Emission Computed Tomography (SPECT) which is substantially more expensive than EBCT and contrary to EBCT, requires the ingestion of radioactive tracing material. To investigate this question we had available data collected at the Radiology Clinic of the Nuclear Imaging Group at Cedars Sinai Medical Center. Asymptomatic patients with high risk factors for coronary artery disease were referred by their doctors to have EBCT performed (Braun et al, 1996) at the clinic. However, of the 5130 males age 36–88 on whom EBCT was performed, only 379 patients were sent by the clinic doctors for assessment of myocardial ischemia. Verification status was significantly associated with daily aspirin use (p value < 0.01). Indeed, we can reasonably suspect that a doctor's decision to send a patient to verification was based on the patient clinical records and not just on the data recorded at the radiology center, i.e. not on the data available to us for analysis. Thus, non-ignorable verification is most likely the case in this problem.

This paper is organized as follows. Section 2 formulates the problem, section 3 discusses identification issues and motivates the assumptions made in this paper to identify the model parameters. Section 4 describes the estimators and Section 5 describes an example.

2 The Study Design

Suppose that on each subject i independently sampled from a population of interest, $i = 1 \dots n$, we measure a continuous or ordinal diagnostic test result (also referred throughout as the “marker”) Y_i and a vector of additional health and demographic variables V_i . Suppose that a gold standard determining the true disease status X_i (X_i being equal to 1 in the presence of disease of interest, and 0 otherwise) is only measured on a subset of the n patients. Let R_i be the indicator that X_i is observed. Assume the patients are a simple random sample of a large population of interest so that (X_i, Y_i, V_i, R_i) , $i = 1 \dots n$ are independent and identically distributed copies of the random vector (X, Y, V, R) .

Under our setting, the observed data are comprised of n iid copies $O_i = (R_i, \alpha(R_i, X_i, Y_i, V_i))$ of the random vector $O = (R, \alpha(R, X, Y, V))$ where $\alpha(1, X, Y, V) = (X, Y, V)$ and $\alpha(0, X, Y, V) = (Y, V)$. Our interest is to estimate the parameter β_0 indexing a parametric model for the conditional density of Y given the covariates V and disease status X ,

$$f(Y|V, X) = f(Y|V, X; \beta_0) \quad (1)$$

where for each β , $f(Y|V, X; \beta)$ is a known conditional density and β_0 is a unknown $p \times 1$ vector.

3 Identification

To help motivate our estimation method and the mathematical results supporting the sensitivity analysis that we advocate in this paper, in this section we will consider nonparametric estimation of $f(Y|V, X)$ under a non-parametric full-data model, that is, under a model that places no restrictions on the joint distribution of (Y, X, V) .

The distribution $f(Y|V, X)$ is not identified by the observed data. This can be easily seen by writing

$$f(Y|V, X=x) = \frac{\left[\sum_{r=0}^1 \Pr(X=x|Y, V, R=r) \Pr(R=r|Y, V) \right] f(Y|V)}{\int \sum_{r=0}^1 \{ \Pr(X=x|y, V, R=r) \Pr(R=r|y, V) \} f(y|V) dy}$$

The expression $\Pr(X=x|Y, V, R=0)$ in the right hand side is not identified by the observed data because it is the probability of disease among non-verified subjects with a given value of the marker and the covariates. It therefore follows that $f(Y|V, X=x)$ is also not identified.

To identify $f(Y|V, X)$ we will assume a pattern-mixture model (Little and Rubin, 2002, Little, 1994, Robins, Rotnitzky, and Scharfstein, 2000) that specifies the connection between distributions $\Pr(X=1|Y, V, R=1)$ and $\Pr(X=1|Y, V, R=0)$. Specifically, we will assume that

$$\Pr(X=1|Y, V, R=0) = \frac{\Pr(X=1|Y, V, R=1)e^{q(Y,V)}}{c(Y, V)} \quad (2)$$

where $q(y, v)$ is an arbitrary, specified, function and

$$c(Y, V) \equiv E \left[e^{q(Y,V)X} | Y, V, R=1 \right] = 1 - \Pr(X=1|Y, V, R=1) \left[1 - e^{q(Y,V)} \right]$$

is a normalizing factor.

The density $f(Y|V, X)$ is identified under model 2. To see this note that since $\Pr(X=1|Y, V, R=1)$ is a functional of the observed data distribution and the right hand side of (2) depends only on $\Pr(X=1|Y, V, R=1)$, then $\Pr(X=1|Y, V, R=0)$ is identified under (2).

It is clear that since any function $q(y, v)$ can be chosen, then $q(y, v)$ is not identified by the observed data. Since the data provide no information to guide the choice of the function $q(y, v)$ then one reasonable analytical strategy, the one advocated in this paper, is to specify a range of plausible functions and repeat the estimation of β_0 , each time regarding a different function $q(y, v)$ fixed and known as a form of sensitivity analysis.

Models that make explicit assumptions about the verification process are referred to as selection models (Little and Rubin, 2002, Little, 1994, Robins, Rotnitzky, and Scharfstein, 2000). The pattern mixture model (2) is indeed also a selection model. Specifically, by Bayes' rule, assumption (2) is equivalent to

$$\log \left\{ \frac{\Pr(R=0|Y, X, V)}{\Pr(R=1|Y, X, V)} \right\} = h(Y, V) + q(Y, V)X \quad (3)$$

where

$$e^{h(Y,V)} = \frac{\Pr(R=0|Y, V)}{\Pr(R=1|Y, V)} \times \frac{1}{c(Y, V)} \quad (4)$$

This alternative expression provides yet another interpretation for $q(Y, V)$. Specifically, $q(Y, V)$ quantifies the dependence on the unobserved disease status of the logarithm of the odds of non-verification after adjustment for the marker and the observed covariates. Thus $q(Y, V)$ is a measure of the degree of the residual association between X and the probability of verification. It reflects the degree of belief by the investigator about the extent of verification bias after accounting for (Y, V) . The choice $q(Y, V) = 0$ for all V and Y corresponds to the assumption that the missing disease status is missing at random, that is

$$\Pr(R=1|Y, X, V) = \Pr(R=1|Y, V) \quad (5)$$

The identifiability considerations discussed so far assumed that the model for the full data did not place restrictions on the distribution of (Y, V, X) . However, in section 2 we indicated that our goal was to estimate $f(Y|V, X)$ under a parametric model for it. Because of this parametric restriction, $q(Y, V)$ may be identified. One could therefore consider estimating $q(Y, V)$, rather than regarding it as fixed and known. The methods described in this paper do not require that $q(Y, V)$ be known. However, we prefer to continue to regard

$q(Y, V)$ as fixed and known and vary it in a sensitivity analysis. Our reason for this is the following. We consider estimation $f(Y|V, X)$ under a parametric model only because with the sample sizes found in practice, estimation of $f(Y|V, X)$ is unfeasible if V is high dimensional due to the curse of dimensionality. However, we would like these dimension reducing considerations to have as little impact as possible on the validity of inferences about $f(Y|V, X)$. Estimation of $q(Y, V)$ under a parametric model for $f(Y|V, X)$ has to necessarily be highly sensitive to misspecification of this parametric model because the identifiability of $q(Y, V)$ depends solely on the distributional shape assumptions about $f(Y|V, X)$. Thus, in order to reduce the impact of model misspecification, we prefer to take the point of view that $q(Y, V)$ should not be estimated, but instead varied in a sensitivity analysis. In any case, for completeness, in section 4.2.1 we explain how our proposed estimators can be extended to accommodate models in which $q(Y, V)$ is assumed unknown and parametrically modelled. We shall denote with $\mathcal{A}(q)$ the model defined by the restrictions (1) and (2) with q fixed and known.

Note that assumption (3) implicitly assumes that all subjects have a positive probability of being selected for verification, regardless of the marker values and covariates. That is, the assumption implies that

$$\Pr(R=1|Y, X, V) > 0 \text{ with probability 1.} \quad (6)$$

This assumption may not always be realistic, especially when extreme values of the marker or covariates provide convincing evidence in favor or against the presence of disease so that clinicians may never send such patients for verification.

4 Estimation of β_0

4.1 The curse of dimensionality

Though β_0 is identified under model $\mathcal{A}(q)$, unfortunately its estimation is unfeasible due to the curse of dimensionality. That this is the case follows from the results stated in Theorem 1 below. Specifically, this Theorem indicates that if a regular asymptotically linear estimator of β_0 existed, its influence function would have to be, up to a multiplicative constant matrix, of the form

$$U(\beta_0, h, \phi, \varphi) = \Delta_\phi(Y, V, X) + \left[\frac{R}{\pi(h)} - 1 \right] [\Delta_\phi(Y, V, X) - \varphi(Y, V)] \quad (7)$$

for some $p \times 1$ vector functions $\phi(Y, V, X)$ and $\varphi(Y, V)$ where

$$\Delta_\phi(Y, V, X) \equiv \phi(Y, V, X) - E_{\beta_0}[\phi(Y, V, X)|V, X]. \quad (8)$$

and $\pi(h) \equiv \{1 + \exp[h(Y, V) + q(Y, V)X]\}^{-1}$. Since by construction $\Delta_\phi(Y, V, X)$ has zero mean, then for $U(\beta_0, h, \phi, \varphi)$ to have mean zero it must be that

$$E \left\{ \left[\frac{R}{\pi(h)} - 1 \right] [\Delta_\phi(Y, V, X) - \varphi(Y, V)] \right\} = 0$$

However, in section 4.3 we show that if $h^\dagger \rightarrow h$ then $U(\beta_0, h^\dagger, \phi, \varphi)$ does not have mean zero under $f(Y|X, V; \beta_0)$ unless $\varphi(Y, V) = \varphi_{DR}(Y, V)$ where

$$\varphi_{DR}(Y, V) \equiv \frac{\sum_{x=0}^1 \Delta_\phi(Y, V, x) \exp\{xq(Y, V)\} \Pr(X=x|R=1, Y, V)}{\sum_{x=0}^1 \exp\{xq(Y, V)\} \Pr(X=x|R=1, Y, V)} \quad (9)$$

This implies that estimation of β_0 under model $\mathcal{A}(q)$ requires preliminary consistent estimation of either $\varphi_{DR}(Y, V)$ (equivalently of $\Pr(X=1|R=1, Y, V)$) or of $h(Y, V)$. When (V, Y) is a vector with two or more continuous components, these functions cannot be well estimated with the moderate sample sizes found in practice essentially because no two units have values of (V, Y) close enough to each other to allow the borrowing of information needed for smoothing. Hence, unrealistically large sample sizes would be required for any estimator of β_0 to have an approximately centered normal sampling distribution with variance small enough to be of substantive use. It follows that for estimation of β_0 we are forced to impose dimension reducing assumptions on either $\Pr(X=1|R=1, Y, V)$ or $h(Y, V)$. In section 4.2 we will consider estimation under parametric models for $h(Y, V)$. In section 4.3 we will describe doubly robust estimators that are consistent and asymptotically normal so long as a model for $\Pr(X=1|R=1, Y, V)$ or a model for $h(Y, V)$ is correctly specified (but not necessarily both).

4.2 Estimation of β_0 under a parametric model for $h(Y, V)$

In this subsection we will consider estimation of β_0 under a model $\mathcal{B}(q)$ defined like $\mathcal{A}(q)$ but with the additional restriction

$$h(y, v) = h(y, v; \gamma_0) \quad (10)$$

where for each (y, v) , $h(y, v; \gamma)$ is a known smooth function of γ , and γ_0 is a $r \times 1$ vector of unknown parameters.

Model $\mathcal{B}(q)$ assumes that $\Pr(R=1|Y, X, V)$ follows the (parametric) logistic regression model

$$\log \left\{ \frac{\Pr(R=0|Y, X, V)}{\Pr(R=1|Y, X, V)} \right\} = h(Y, V; \gamma_0) + q(Y, V)X \quad (11)$$

It is important to note that model $\mathcal{B}(q)$ is not a parametric model for the joint distribution of (Y, X, V) since it leaves the distributions $f(X|V)$ and $f(V)$ unrestricted. Inference under model $\mathcal{B}(q)$, as opposed to likelihood based inference under a model that additionally imposes parametric restrictions on $f(X|V)$ and $f(V)$, has the advantage of being valid regardless of what the true distributions $f(X|V)$ and $f(V)$ are. In contrast, likelihood based inference under a fully parametric model is generally sensitive to misspecification of the model for $f(X|V)$, thus violating the essential property of regression estimators with full data, of being consistent regardless of what the distribution of the covariates is. Specifically, suppose that models $f(V; \eta_1)$ and $f(X|V; \eta_2)$ for $f(V)$ and $f(X|V)$ indexed by finite dimensional parameters η_1 and η_2 were assumed. In such case, the likelihood contribution of each patient would be

$$L(\beta, \gamma, \eta) = f(V; \eta_1) \left\{ \sum_{x=0}^1 [1 - \Pr(R=1|Y, X=x, V; \gamma)] f(Y|X=x, V; \beta) \Pr(X=x|V; \eta_2) \right\}^{1-R} \\ \times \{\Pr(R=1|Y, X, V; \gamma) f(Y|X, V; \beta) \Pr(X|V; \eta_2)\}^R \quad (12)$$

The term $f(V; \eta_1)$ factors out of the likelihood, but $\Pr(X|V; \eta_2)$ doesn't. Thus, consistent maximum likelihood estimation of β_0 would require correctly specified models for $f(X|V; \eta_2)$ in addition to correct specification of the distribution $\Pr(R=1|Y, X, V; \gamma)$. Maximum likelihood about β is therefore sensitive to misspecification of both models.

Semiparametric model $\mathcal{B}(q)$ is defined by a semiparametric model for the law of the full data (Y, X, V) and a parametric non-ignorable model for the missingness probabilities. A general theory for inference in these models has been derived by Rotnitzky and Robins, 1997. In particular, these authors derived a general representation for the set of all influence functions for regular asymptotically linear estimators of finite dimensional parameters (β, γ) where, as in our problem, β is a parameter that depends on the full-data distribution and γ indexes a parametric model for the missingness probabilities. The following Theorem, specializes the results in Rotnitzky and Robins to obtain the representation of the set orthogonal to the nuisance tangent space for (β, γ) in model $\mathcal{B}(q)$. We subsequently use this representation to derive a class of estimating equations whose solutions are, up to asymptotic equivalence, all regular asymptotically linear estimators of (β_0, γ_0) .

Theorem 1—in model $\mathcal{B}^*(q)$ defined like model $\mathcal{B}(q)$ but with the additional restriction

$$\Pr(R=1|Y, X, V) > \sigma > 0 \quad (13)$$

for some $\sigma > 0$, the orthogonal complement to the nuisance tangent space for (β, γ) is equal to

$$\Lambda_{\text{nuis}}^\perp = \left\{ U(\beta_0, \gamma_0; \phi, \varphi) \equiv \frac{R}{\pi(\gamma_0)} \Delta_\phi(Y, V, X) - A(\gamma_0; \varphi) : E[\phi(Y, V, X)^2] < \infty, E[\varphi(Y, V)^2] < \infty \right\}$$

where for any γ

$$\begin{aligned} \pi(\gamma) &\equiv \Pr(R=1|Y, X, V; \gamma) \equiv [1 + \exp\{h(Y, V; \gamma) + q(Y, V)X\}]^{-1} \\ A(\gamma; \varphi) &\equiv \left[\frac{R}{\pi(\gamma)} - 1 \right] \varphi(Y, V) \end{aligned} \quad (14)$$

and $\Delta_\phi(Y, V, X)$ is defined as in (8).

It is a well known result from semiparametric theory that the influence functions of RAL estimators of (β_0, γ_0) are elements of the orthogonal complement to the nuisance tangent space. It follows from the previous theorem that if a RAL estimator of (β_0, γ_0) exists, then it must be asymptotically equivalent to a solution $(\hat{\beta}(\phi, \varphi), \hat{\gamma}(\phi, \varphi))$ to an estimating equation of the form

$$E_n[U(\beta, \gamma; \phi, \varphi)] = 0 \quad (15)$$

where $\phi(Y, V, X)$ and $\varphi(Y, V)$ are now random vectors with dimension $(p+r) \times 1$ and throughout we use the notational convention that for arbitrary random vectors

$B_1, \dots, B_n, E_n(B) \equiv n^{-1} \sum_{i=1}^n B_i$. This is so because, under regularity conditions, using standard Taylor expansion arguments it can be shown that

$$\sqrt{n} \left\{ \left(\widehat{\beta}(\phi, \varphi), \widehat{\gamma}(\phi, \varphi) \right)' - (\beta_0, \gamma_0)' \right\} = \frac{1}{\sqrt{n}} \Gamma(\beta_0, \gamma_0; \phi, \varphi)^{-1} \sum_{i=1}^n U_i(\beta_0, \gamma_0; \phi, \varphi) + o_p(1)$$

where

$$\Gamma(\beta_0, \gamma_0; \phi, \varphi) = \frac{\partial}{\partial(\beta, \gamma)} E[U(\beta, \gamma; \phi, \varphi)] \Big|_{(\beta, \gamma) = (\beta_0, \gamma_0)}$$

A heuristic motivation for the class of estimating equations (15) is as follows. Consider first estimation of β_0 when X is always observed. Any RAL estimator of β_0 has to be asymptotically equivalent to an estimating equation of the form

$$E_n\{\Delta_\phi(Y, V, X; \beta)\} = 0 \quad (16)$$

where

$$\Delta_\phi(Y, V, X; \beta) \equiv \phi(Y, X, V) - \int \phi(y, X, V) f(y|X, V; \beta) dy$$

That this must be the case follows from the fact, that any mean zero function of (Y, X, V) that is orthogonal to nuisance scores in the full data model must be of the form $\Delta_\phi(Y, V, X)$. The proof of this result as well as the proofs of all the theorems in this paper can be found in the technical report Page and Rotnitzky (2008) available from the authors upon request.

When, as in our problem, X is not observed for some subjects and the response probabilities depend on Y we can no longer use the estimating equation (16) since subjects with observed X are a biased sample, and for them $\Delta_\phi(Y, V, X)$ no longer has mean zero.

If the response probabilities $\pi(\gamma_0)$ were known, we could estimate β solving the inverse probability weighted estimating equation

$$E_n \left\{ \frac{R}{\pi(\gamma_0)} \Delta_\phi(Y, V, X; \beta) \right\} = 0$$

This equation is unbiased because

$$E \left\{ \frac{R}{\pi(\gamma_0)} \Delta_\phi(Y, V, X; \beta_0) \right\} = E \left\{ \frac{E(R|Y, V, X)}{\pi(\gamma_0)} \Delta_\phi(Y, V, X; \beta_0) \right\} = 0$$

Since γ_0 is unknown we must replace it by a consistent estimator of it. However, we can't estimate γ_0 fitting a logistic regression model of R on X, V and Y with offset $q(Y, V)$ because X is missing when $R = 0$. Instead, we can solve the estimating equation

$$E_n[A(\gamma, \varphi_\gamma)] = 0 \quad (17)$$

where φ_γ is an arbitrary, user specified, $r \times 1$ vector function. Note that $A(\gamma, \varphi_\gamma)$ depends only on the observed data and thus is a computable quantity. Since $A(\gamma_0, \varphi_\gamma)$ has mean zero for any function $\varphi_\gamma(Y, V)$, then under standard regularity conditions, and provided (13) is true, the estimating equation (17) has a solution $\hat{\gamma}$ that is a consistent and asymptotically normal estimator of γ_0 . Now, under standard regularity conditions, the equation

$$E_n \left\{ \frac{R}{\pi(\gamma)} \Delta_\phi(Y, V, X; \beta) \right\} = 0 \quad (18)$$

will have a solution $\hat{\beta}_w$ that will be consistent and asymptotically normal for estimating β_0 .

The solution $\hat{\beta}_w$ is not entirely satisfactory because it may be very inefficient. Specifically, data from subjects with X missing enters into the estimation of β only through γ . It is indeed possible to find a larger class of estimating equations with a member which has a solution whose limiting distribution has a smaller asymptotic variance than that of $\hat{\beta}_w$. Specifically,

since $A(\gamma_0, \varphi)$ and $\frac{R}{\pi(\gamma_0)} \Delta_\phi(Y, V, X; \beta_0)$ both have mean zero under (γ_0, β_0) we can consider

concatenating terms of the form $\frac{R}{\pi(\gamma)} \Delta_\phi(Y, V, X; \beta) - A(\gamma, \varphi)$ thus creating joint estimating equations of the form (15) for (γ_0, β_0) . Theorem 1 establishes that the solutions of estimating equations of this form comprise, up to asymptotic equivalence, all possible unbiased estimating equations for (γ_0, β_0) .

4.2.1 Estimation of the selection bias function $q(Y, V)$: Suppose that instead of regarding $q(Y, V)$ as known, we assume a parametric model

$$q(Y, V) = q(Y, V; \xi_0)$$

where for each (Y, V) , $q(Y, V; \xi)$ is a smooth function of ξ and ξ_0 is an unknown $t \times 1$ parameter vector. Theorem 1 remains true with the following modifications:

$$\Lambda_{\text{nuis}}^\perp = \left\{ U(\beta_0, \gamma_0, \xi_0; \phi, \varphi) \equiv \frac{R \Delta_\phi(Y, V, X)}{\pi(\gamma_0, \xi_0)} - A(\gamma_0, \xi_0; \varphi) : E[\phi(Y, V, X)^2] < \infty, E[\varphi(Y, V)^2] < \infty \right\}$$

where for any γ, ξ ,

$$\begin{aligned} \pi(\gamma, \xi) &\equiv \Pr(R=1|Y, X, V; \gamma, \xi) \equiv [1 + \exp\{h(Y, V; \gamma) + q(Y, V; \xi)X\}]^{-1} \\ A(\gamma, \xi; \varphi) &\equiv \left[\frac{R}{\pi(\gamma, \xi)} - 1 \right] \varphi(Y, V) \end{aligned}$$

Consequently, if a RAL estimator of (β, γ, ξ) exists, it will be asymptotically equivalent to a solution $(\hat{\beta}(\phi, \varphi), \hat{\gamma}(\phi, \varphi), \hat{\xi}(\phi, \varphi))$ to an estimating equation of the form

$$E_n[U(\beta, \gamma, \xi; \phi, \varphi)] = 0$$

where $\phi(Y, V, X)$ and $\phi(Y, V)$ are now random vectors with dimension $(p + r + t) \times 1$.

4.3 Doubly Robust Estimation

The estimators $\hat{\beta}(\phi, \varphi)$ of the previous subsection can be severely biased if the model (10) is misspecified because some verified subjects may receive inappropriate weight. Interestingly, there is a specific choice of $\phi(Y, V)$ that gives some protection to misspecification of (10).

Specifically, consider the function $\varphi_{DR}(Y, V)$ defined in (9) which can be written as

$$\varphi_{DR}(Y, V) = \frac{E[e^{q(Y,V)X} \Delta_\phi(Y, X, V) | R=1, Y, V]}{E[e^{q(Y,V)X} | R=1, Y, V]} \quad (19)$$

$$= E[\Delta_\phi(Y, X, V) | R=0, Y, V] \quad (20)$$

where the last equality is true under (2). Remarkably, it can be shown that $U(\beta_0, \gamma^*; \phi, \varphi_{DR})$ has mean zero even if γ^* is different from γ_0 ; that is,

$$E[U(\beta_0, \gamma^*; \phi, \varphi_{DR})] = 0 \quad (21)$$

Equation (21) implies that if $\hat{\varphi}_{DR}(Y, V)$ were a consistent estimator of $\varphi_{DR}(Y, V)$ converging at a sufficiently fast rate, then the solution $\hat{\beta}$ of

$$E_n \left\{ \frac{R}{\pi(\gamma)} \Delta_\phi(Y, V, X) - \left[\frac{R}{\pi(\gamma)} - 1 \right] \hat{\varphi}_{DR}(Y, V) \right\} = 0$$

where $\phi(Y, V, X)$ is an arbitrary $p \times 1$ function of (Y, V, X) and $\hat{\gamma}$ solves

$$E_n \{A(\gamma, \varphi_\gamma)\} = 0$$

for an arbitrary $r \times 1$ function $\varphi_\gamma(Y, V)$, would be a consistent and asymptotically normal estimator of β_0 even if model (10) is misspecified. Unfortunately, when (Y, V) has two or more continuous components we can't hope to estimate $\varphi_{DR}(Y, V)$ well using smoothing techniques with the moderate sample sizes found in practice due to the curse of dimensionality. We can, however, consider estimating $\varphi_{DR}(Y, V)$ under a parametric model for it. To pose such model we note that, in view of the definition of $\varphi_{DR}(Y, V)$, it suffices to postulate a parametric model for $\Pr(X=1 | R=1, Y, V)$. That is,

$$\Pr(X=1 | R=1, Y, V) = m(Y, V; \mu_0) \quad (22)$$

where for each y, v , $m(y, v; \mu)$ is a smooth function of μ and μ_0 is an unknown $I \times 1$ parameter vector.

A consistent estimator $\hat{\mu}$ of μ_0 under model (22) can be obtained by solving unbiased estimating equations for binary regression (based on the subjects with observed X 's) of the form

$$E_n\{H(\mu;d)\}=0 \quad (23)$$

where $H(\mu; d) \equiv Rd(Y, V) \{X - m(Y, V; \mu)\}$ and $d(Y, V)$ is a user specified $l \times 1$ function of Y and V .

Define $\varphi_{DR}(Y, V; \mu)$ like $\varphi_{DR}(Y, V)$ but with $m(Y, V; \mu)$ replacing $\Pr(X=1|R=1, Y, V)$ and in a slight abuse of notation write

$$U(\beta, \gamma, \mu, \phi) \equiv \frac{R}{\pi(\gamma)} \Delta_\phi(Y, V, X; \beta) - \left[\frac{R}{\pi(\gamma)} - 1 \right] \varphi_{DR}(Y, V; \mu). \text{ Let}$$

$$M(\beta, \gamma, \mu, \phi, \varphi_\gamma, d) \equiv \begin{bmatrix} U(\beta, \gamma, \mu, \phi) \\ A(\gamma; \varphi_\gamma) \\ H(\mu; d) \end{bmatrix}$$

where $\phi(Y, V, X)$ and $\varphi_\gamma(Y, V)$ are arbitrary $p \times 1$ and $r \times 1$ vector functions and $A(\gamma; \varphi_\gamma)$ is defined as in (14) but with φ_γ replacing ϕ .

The following Theorem establishes the asymptotic distribution of an estimator $\hat{\beta}_{DR}(\phi, \varphi_\gamma, d)$ of β_0 which solves, together with $\hat{\gamma}$ and $\hat{\mu}$, the estimating equation

$$E_n\{M(\beta, \gamma, \mu, \phi, \varphi_\gamma, d)\}=0 \quad (24)$$

The Theorem establishes that the estimator $\hat{\beta}_{DR}(\phi, \varphi_\gamma, d)$ is a doubly robust consistent and asymptotically normal (CAN) estimator of β_0 in the sense that it is CAN if one of models (10) or model (22), but not necessarily both, is correctly specified. The Theorem uses the following definitions. γ^* and μ^* denote the probability limits of $\hat{\gamma}$ and $\hat{\mu}$ respectively, $U \equiv U(\beta_0, \gamma^*, \mu^*, \phi)$,

$$\begin{aligned} \Upsilon &\equiv E \left[\frac{\partial U(\beta_0, \gamma, \mu^*; \phi)}{\partial \gamma} \Big|_{\gamma=\gamma^*} \right] \left\{ E \left[\frac{\partial A(\gamma; \varphi_\gamma)}{\partial \gamma'} \Big|_{\gamma=\gamma^*} \right] \right\}^{-1} \times A(\gamma^*; \varphi_{\gamma^*}) \\ \Psi &\equiv E \left[\frac{\partial U(\beta_0, \gamma^*, \mu; \phi)}{\partial \mu} \Big|_{\mu=\mu^*} \right] \left\{ E \left[\frac{\partial H(\mu; d)}{\partial \mu'} \Big|_{\mu=\mu^*} \right] \right\}^{-1} \times H(\mu^*; d) \end{aligned}$$

and

$$\Sigma = E \left[\frac{\partial U(\beta, \gamma^*, \mu^*; \phi)}{\partial \beta'} \Big|_{\beta=\beta_0} \right]$$

Theorem 2—Suppose model $A(q)$ holds. Then under regularity conditions, if either (10) and (13) hold or (22) holds,

1. the estimating equation (24) has a solution $(\hat{\beta}_{DR}(\phi, \varphi_\gamma, d), \hat{\gamma}, \hat{\mu})$ which satisfies

$$\sqrt{n} \{ \hat{\beta}_{DR}(\phi, \varphi_\gamma, d) - \beta_0 \} = \sum^{-1} \sqrt{n} E_n \{ U - \Upsilon - \Psi \} + o_p(1)$$

Thus, $\sqrt{n} \{ \hat{\beta}_{DR}(\phi, \varphi_\gamma, d) - \beta_0 \} \rightarrow N \left(0, \sum^{-1} \Gamma \sum^{-1'} \right)$ where $\Gamma = \text{var}(U - \Upsilon - \Psi)$.

2. A consistent estimator of $\Sigma^{-1} \Gamma \Sigma^{-1'}$ is given by $\hat{\Sigma}^{-1} \hat{\Gamma} \hat{\Sigma}^{-1'}$ where

$$\hat{\Sigma} = E_n \left[\frac{\partial U(\beta, \gamma, \mu; \phi)}{\partial \beta'} \right]_{\beta = \hat{\beta}_{DR}}, \quad \hat{\Gamma} = E_n \left((\hat{U} - \hat{\Upsilon} - \hat{\Psi}) (\hat{U} - \hat{\Upsilon} - \hat{\Psi})' \right) \text{ with } (\hat{\beta}_{DR} \equiv \hat{\beta}_{DR}(\phi, \varphi_\gamma, d), \hat{U} \equiv U(\hat{\beta}_{DR}, \gamma, \mu, \phi),$$

$$\hat{\Upsilon} \equiv E_n \left[\frac{\partial U(\hat{\beta}_{DR}, \gamma, \mu; \phi)}{\partial \gamma} \right]_{\gamma = \hat{\gamma}} \left\{ E_n \left[\frac{\partial A(\gamma; \varphi_\gamma)}{\partial \gamma'} \right]_{\gamma = \hat{\gamma}} \right\}^{-1} \times A(\hat{\gamma}; \varphi_\gamma)$$

and

$$\hat{\Psi} \equiv E_n \left[\frac{\partial U(\hat{\beta}_{DR}, \gamma, \mu; \phi)}{\partial \mu} \right]_{\mu = \hat{\mu}} \left\{ E_n \left[\frac{\partial H(\mu; d)}{\partial \mu'} \right]_{\mu = \hat{\mu}} \right\}^{-1} \times H(\hat{\mu}; d)$$

4.3.1 Further remarks: The asymptotic variance of $\hat{\beta}_{DR}(\phi, \varphi_\gamma, d)$ depends on the user's choice of functions ϕ , φ_γ and d . Therefore, it would be desirable to find functions ϕ_{opt} , $\varphi_{\gamma, opt}$ and d_{opt} that are optimal in the sense that $\hat{\beta}_{DR}(\phi_{opt}, \varphi_{\gamma, opt}, d_{opt})$ has the smallest variance (in the positive definite sense) of the asymptotic variances of all $\hat{\beta}_{DR}(\phi, \varphi_\gamma, d)$. Unfortunately, it can be shown that the optimal functions solve an exceedingly difficult integral equation. Thus, we forego trying to obtain ϕ_{opt} , $\varphi_{\gamma, opt}$ and d_{opt} . Nevertheless, following the results of Hansen (1982), and Qu, Lindsay and Li (2000), one can obtain estimators with good efficiency properties by solving an estimating equation based on an optimal linear combination of a finite number of unbiased doubly-robust estimating functions.

Our doubly-robust estimator of β_0 requires that we specify parametric models (22) and (11) for $\Pr(X=1|R=1, Y, V)$ and $\Pr(R=1|Y, X, V)$ respectively. However, these models may be incompatible in the sense that the restrictions they imply on the distribution $f(Y|X, V)$ may never hold simultaneously. This apparent contradiction does not pose any practical difficulty. By posing two models, (22) and (11), we certainly obtain better protection against model misspecification than posing just model (11). The fact that (22) and (11) cannot hold simultaneously is inconsequential since our doubly-robust estimators are consistent and asymptotically normal so long as one, but not necessarily both models is correct.

On occasion, one might have available a vector containing a large number of variables V thought to be simultaneous predictors of verification and disease status. However, one might want to estimate a model for the distribution of the marker conditional only on a function $Z = b(V)$ of V . As an example, in the application illustrated in section 5, we had available records on age and whether or not the patient was using aspirin daily. However, we were not interested in a refined assessment of how the predictive ability of the marker changed with age, primarily because we had insufficient a-priori information as to how to model this association adequately. Instead, we wanted to evaluate if, quite generally, the marker's predictive accuracy was better in the elderly. As such, we considered $V = (V_1, V_2)$, where $V_1 = \text{age}$ and $V_2 = \text{indicator of aspirin use}$, and $Z = (I(V_1 \geq 55), V_2)$. The results in this paper, including the double-robustness of $\hat{\beta}_{DR}(\phi, \varphi_\gamma, d)$ remain valid verbatim if V is replaced by Z in every expression where it appears as a conditioning variable for the distribution of Y .

5 Example

We applied the above methodology to the dataset described in the introduction. Electron-beam computerized tomography (EBCT) of the coronary arteries is a method for measuring coronary artery calcium, and is an established marker of atherosclerotic plaque. It is unknown however, whether EBCT is an effective screening tool for myocardial ischemia. This is an important question because the other available noninvasive instrument for screening myocardial ischemia is the Dual Isotope Myocardial Perfusion Single Photon Emission Computed Tomography (SPECT) which is substantially more expensive than EBCT and contrary to EBCT, requires the ingestion of radioactive tracing material.

Of the 5130 males age 36–88 on whom EBCT was performed at the radiology clinic, 379 patients were sent by the clinic doctors for assessment of myocardial ischemia with SPECT, and of them 28 were found positive. EBCT histograms stratified by verification and disease status (figures not shown), suggest that compared to the EBCT distribution of the non-verified subjects, the distributions of verified subjects are markedly skewed to the right indicating that subjects with lower marker values are less likely than others to be verified.

Here we consider estimation of the distribution of $Y = \log(EBCT + 1)$ conditional on the prognostic factors age and use of daily aspirin. We had available records on the variable $V = (V_1, V_2)$ with V_1 age and V_2 the binary indicator of daily use of aspirin, but for reasons indicated in section 4.3.3 we were interested in fitting a model for $f(Y|Z, X)$ where $Z = (Z_1, Z_2)$, $Z_1 = I(V_1 \leq 55)$, $Z_2 = V_2$ and X is the indicator of coronary disease as determined by SPECT. Our model (1) for the conditional distribution of Y given covariates Z and disease status was

$$f(Y|X, Z; \beta) = p(X, Z; \beta_p)^{I(Y=0)} \left[\frac{1 - p(X, Z; \beta_p)}{\sqrt{2\pi\sigma^2(X, Z; \beta_v)}} e^{\frac{(Y - \omega(X, Z; \beta_m))^2}{2\sigma^2(X, Z; \beta_v)}} \right]^{I(Y \neq 0)}$$

where $p(X, Z; \beta_p) = (1 - X)(1 + e^{-D'\beta_p})^{-1}$, $\omega(X, Z; \beta_m) = W'\beta_m$, $\sigma^2(X, Z; \beta_v) = e^{-B'\beta_v}$, $W = (X, Z_1, Z_2, 1)'$, $B = (1, X)'$ and $D = (Z_1, Z_2, 1)'$. This model assumes that the conditional distribution of Y has a discrete mass probability equal to $p(X, Z; \beta_p)$ at $Y = 0$ and is normal with conditional mean $\omega(X, Z; \beta_m)$ and conditional variance $\sigma^2(X, Z; \beta_v)$ otherwise. Note that in particular, our model assumes that no disease patient has marker values equal to zero.

Our estimating equations used

$$\phi(Y, Z, X; \beta) = \begin{bmatrix} W(Y - W'\beta_m)e^{-B'\beta_v}I(Y>0) \\ B \left\{ e^{-B'\beta_v}(Y - W'\beta_m)^2 - 1 \right\} I(Y>0) \\ DI(Y=0)I(X=0) \end{bmatrix}$$

In our analysis we estimated $\beta = (\beta_m, \beta_v, \beta_p)$ using $q(Y, V) = \tau$, repeating the analysis for values of τ equal to -3 , 0 and 1 . These values were chosen so as to include settings with: 1) very severe residual selection bias in favor of diseased subjects (the choice $\tau = -3$ indicates that the odds of selection is $\exp(3) \approx 20$ times larger for diseased than healthy patients with the same values of EBCT, age and aspirin use), 2) ignorable verification ($\tau = 0$), and 3) the less likely nonnegligible residual selection bias in favor of the healthy subjects ($\tau = 1$). We

used working selection and disease models with $h(Y, V; \gamma) = \gamma_0 + \gamma_1 I(Y=0) + \gamma_2 I(Y>0) (Y-3) + \gamma_3 I(Y>0) V_2$ and $m(Y, V; \mu) = (1 + \exp(-\mu_0 - \mu_1 Y - \mu_2 V_1 - \mu_3 V_2))^{-1}$. The model for $m(Y, V)$ was chosen after using standard model selection techniques for logistic regression. In the model for $h(Y, V)$ we included a separate intercept for subjects with $Y=0$ so that in computing $\hat{\beta}_{DR}$ the 13 verified subjects with marker value equal to 0 would be given weights that would make them represent the 1749 non-verified subjects with zero marker value. The function $\phi_\gamma(Y, V_2)$ used in the estimating functions $A(\gamma, \phi_\gamma)$ was chosen to be equal to the vector of regressors in the assumed selection model. Similarly, the function $d(Y, V_2, V_3)$ used in the estimating function $H(\mu; d)$ was chosen to be equal to the vector of covariates in the assumed model for disease.

Figures 1–4 display the estimated log-likelihood ratio function $\log \{f(y|Z=z, X=1; \hat{\beta}) / f(y|Z=z, X=0; \hat{\beta})\}$ for values of z representing the four sub-groups defined by age and aspirin use status and for values of y in the range of observed values of Y . The figure displays the estimated log-likelihood ratio using the doubly-robust estimates of β for the choices $q = -3$, $q = 0$ and $q = 1$. Overall, lower values of the marker appear effective at ruling out myocardial ischemia, however higher values of the marker do not appear particularly useful at ruling in disease. The marker has noticeably better ability for ruling out myocardial ischemia in the elderly, but it appears otherwise similar in the two aspirin groups. The marker's ability for ruling in ischemia at higher values of the marker is similarly poor in the four groups.

For the most part the log-likelihood ratio estimated using only verified subjects is lower across all values of y than those based on adjustment for selection bias regardless of the value of q , indicating that an analysis based on verified only subjects will overestimate specificity but underestimate sensitivity. Interestingly, among those who were users of daily aspirin aged greater than 55 (the group most likely to have had their disease status verified) all the estimators resulted in estimates for the log-likelihood ratio that were not very different from each other. The results suggest that in settings with very severe residual selection bias in favour of diseased subjects ($q = -3$), the estimators assuming no residual selection bias (i.e. $q = 0$) would in general overestimate the likelihood ratio at lower values of the marker.

Acknowledgments

John Page was supported by grant # 0475016N of the American Heart Association and Andrea Rotnitzky was supported by NIH grants R29 GM48704 and R01 AI32475.

References

1. Baker SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics*. 1995; 51:330–337. [PubMed: 7539300]
2. Braun J, Oldendorf M, et al. Electron beam computed tomography in the evaluation of cardiac calcifications in chronic dialysis patients. *Am J Kidney Dis*. 1996; 27:394–401. [PubMed: 8604709]
3. Gray R., Begg CB, Greenes RA. Construction of the receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making*. 1984; 4:151–164. [PubMed: 6472063]
4. Hansen LP. Large sample properties of generalized methods of moments estimators. *Econometrica*. 1982; 50:1029–1054.
5. Kosinski AS, Barnhart HX. Accounting for Non-ignorable Verification Bias in Assessment of Diagnostic Test. *Biometrics*. 2003; 59(2)
6. Little R. A class of pattern-mixture models for normal missing data. *Biometrika*. 1994; 81:471–483.
7. Little, R.; Rubin, D. *Statistical Analysis with Missing Data*. J. Wiley; N. York: 2002.

8. Page, J.; Rotnitzky, A. Technical Report. Harvard School of Public Health, Dep. of Biostatistics; 2008. Proofs of the assertions in the paper Estimation of the disease-specific diagnostic marker distribution under verification bias, by Page and Rotnitzky.
9. Qu A, Lindsay B, Li B. Improving generalized estimating equations using quadratic inference functions. *Biometrika*. 2000; 87:823–836.
10. Robins, JM.; Rotnitzky, A.; Scharfstein, DO. Sensitivity analysis for selection bias and unmeasured confounding in missing data and causal inference models. In: Halloran, E.; Berry, D., editors. *Statistical Models for Epidemiology, the Environment, and Clinical Trials*. 2000. p. 1-95.
11. Rodenberg C, Zhou XH. ROC curve estimation when covariates affect the verification process. *Biometrics*. 2000; 56:131–136.
12. Rotnitzky A, Robins J. Analysis of semi-parametric regression models with non-ignorable non-response. *Statistics in Medicine*. 1997; 16:81–102. [PubMed: 9004385]
13. Rotnitzky A, Faraggi D, Schisterman E. Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. To appear. *Journal of the American Statistical Association Theory and Methods*. 2005
14. Rubin D. Inference and missing data. *Biometrika*. 1976; 72:581–92.
15. Toledano A, Gatsonis C. Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine*. 1996; 15:1807–26. [PubMed: 8870162]
16. Toledano A, Gatsonis C. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*. 1999; 55:488–96. [PubMed: 11318205]
17. Zhou X. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics - Theory and Methods*. 1993; 54:124–135.
18. Zhou XH. Effect of verification bias on positive and negative predictive values. *Statistics in Medicine*. 1994; 13:1737–1745. [PubMed: 7997707]
19. Zhou XH. Nonparametric ML estimate of a ROC area corrected for verification bias. *Biometrics*. 1996; 52:310–316.
20. 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]
21. Zhou XH. Comparing the correlated areas under the ROC curves of two diagnostic tests in the presence of verification bias. *Biometrics*. 1998b; 54:453–470. [PubMed: 9629639]
22. Zhou XH, Rodenberg C. Estimating the ROC curve in the presence of nonignorable verification bias. *Commun in Statistics, Theory and Methods*. 1998; 27:635–57.
23. Zhou XH, Higgs R. Assessing the relative accuracies of two screening tests in the presence of verification bias. *Statistics in Medicine*. 2000; 19:1697–1705. [PubMed: 10844728]
24. Zhou, XH.; Obuchowski, NA.; McClish. *Statistical Methods in Diagnostic Medicine*. Wiley; New York: 2002.
25. Zhou XH, Castelluccio P. Nonparametric analysis for the ROC areas of two diagnostic tests in the presence of nonignorable verification bias. *Journal of Statistical Planning and Inference*. 2003; 115:193–213.

Figure 1.
Sensitivity Analysis of the Diagnostic Likelihood Ratio
Age < 55. non-aspirin users

Figure 2.
Sensitivity Analysis of the Diagnostic Likelihood Ratio
Age < 55. Daily aspirin users

Figure 3.
Sensitivity Analysis of the Diagnostic Likelihood Ratio
Age ≥ 55 . non-aspirin users

Figure 4.
Sensitivity Analysis of the Diagnostic Likelihood Ratio
Age ≥ 55 . Daily aspirin users