



Published in final edited form as:

Stat Med. 2009 February 1; 28(3): 361–376. doi:10.1002/sim.3388.

Direct Estimation of the Area Under the Receiver Operating Characteristic Curve in the Presence of Verification Bias

Hua He^{1,†}, Jeffrey M. Lyness^{2,†}, and Michael P. McDermott^{1,*}

¹ Department of Biostatistics and Computational Biology, University of Rochester Medical Center, 601 Elmwood Ave, Box 630, Rochester, NY 14642, U.S.A

² Department of Psychiatry, Geriatric Psychiatry Program, University of Rochester Medical Center, 601 Elmwood Ave, Box PSYCH, Rochester, NY 14642, U.S.A

SUMMARY

The area under a receiver operating characteristic (ROC) curve (AUC) is a commonly used index for summarizing the ability of a continuous diagnostic test to discriminate between healthy and diseased subjects. If all subjects have their true disease status verified, one can directly estimate the AUC nonparametrically using the Wilcoxon statistic. In some studies, verification of the true disease status is performed only for a subset of subjects, possibly depending on the result of the diagnostic test and other characteristics of the subjects. Because estimators of the AUC based only on verified subjects are typically biased, it is common to estimate the AUC from a bias-corrected ROC curve. The variance of the estimator, however, does not have a closed-form expression and thus resampling techniques are used to obtain an estimate. In this paper, we develop a new method for directly estimating the AUC in the setting of verification bias based on U-statistics and inverse probability weighting. Closed-form expressions for the estimator and its variance are derived. We also show that the new estimator is equivalent to the empirical AUC derived from the bias-corrected ROC curve arising from the inverse probability weighting approach.

Keywords

Diagnostic test; Inverse probability weighting; Missing at random; U-statistic

1 INTRODUCTION

For a diagnostic test that yields a continuous test result, the receiver operating characteristic (ROC) curve is a popular tool for displaying the ability of the test to discriminate between healthy and diseased subjects. The continuous test result can be dichotomized at a specified cutpoint and the sensitivity and specificity can be computed. When one varies the cutpoint throughout the entire real line, the resulting pairs (1 – specificity, sensitivity) form the ROC curve. The area under the ROC curve (AUC) is commonly used as a summary index of the accuracy of the diagnostic test. The AUC can be interpreted as $\Pr(T_1 > T_2)$, where T_1 is the test result from a randomly selected diseased subject and T_2 is the test result from a randomly selected non-diseased subject [1,2].

In the complete data case where the true disease status for every subject is verified using a gold standard evaluation, the sensitivity and specificity at all possible cutpoints can be easily

*Correspondence to: Michael P. McDermott, Department of Biostatistics and Computational Biology, University of Rochester, 601 Elmwood Ave, Box 630, Rochester, NY 14642, U.S.A.

†Email: huahe@bst.rochester.edu, Jeffrey Lyness@urmc.rochester.edu, mikem@bst.rochester.edu

estimated by simple proportions. The nonparametric (empirical) estimate of the ROC curve can then be obtained, and the area under the empirical ROC curve is a natural estimate of the AUC. Equivalently, one can directly estimate the AUC using the nonparametric Wilcoxon statistic [1,2].

In many situations, not all subjects given the new diagnostic test ultimately have their true disease status verified. There are various reasons for this. For example, some gold standard evaluations are expensive and time consuming, and some are based on invasive procedures such as surgery. In these situations, subjects with negative test results may be less likely to undergo a gold standard evaluation than subjects with positive test results. When the decision regarding whether or not to verify the subject's true disease status depends on the test result (and possibly other subject characteristics), estimators of AUC based only on data from the verified subjects may be badly biased [3–6]. This is called verification bias [3] or work-up bias [7].

Most existing methods for correcting verification bias are applicable only for tests that yield binary or ordinal results [3,8–17]. With few exceptions [10,12,13,16,17] these methods assume that the true disease status, if missing, is missing at random (MAR) [18], i.e., that the probability of a subject having the disease status verified is purely determined by the test result and the subject's observed characteristics, and is conditionally independent of the unknown true disease status. Two recent papers have considered methods for bias correction when the test result is continuous [19,20]. The examples cited in these papers clearly illustrate the need for the development of methods for this case. Alonzo and Pepe [19] extended Begg and Greenes' approach [3] from a binary test with categorical covariates to a continuous test with continuous covariates. Alonzo and Pepe [19] also recognized that a study with verification-biased sampling can be thought of as a study with a two-phase or double-sampling design; a good review of methodology for prevalence estimation under these designs is provided by Carroll et al. [21]. Alonzo and Pepe [19] applied several of these methods to the problem of correcting verification bias, including the inverse probability weighting (IPW) approach [22], the mean score (MS) method [23,24], and a semi-parametric efficient approach [25,26]. These methods can be used to estimate a bias-corrected sensitivity and specificity pair at each possible cutpoint and an empirical bias-corrected ROC curve can then be constructed. A bias-corrected estimate of AUC can be derived as the area under the empirical bias-corrected ROC curve. There is no closed-form expression for the variance of the AUC estimator; therefore resampling methods, such as the bootstrap, are frequently used for inference.

Rotnitzky et al. [20] proposed a doubly robust estimator of the AUC. The estimator requires specification of parametric models for the probability of disease and the probability of verification of disease status, but it is consistent and asymptotically normal if either (not necessarily both) of these models is correctly specified. The estimator can also be applied in the situation where the missingness mechanism for the true disease status is nonignorable.

In this paper we develop a new method for directly estimating the AUC in the presence of verification bias when the test result is continuous. We derive a closed-form expression for its asymptotic variance. The new estimator is particularly useful in cases where the mechanism for deciding whether or not to verify the subject's true disease status is well understood or can be controlled by the investigators. In Section 2, we give a brief review of existing methods for estimating the AUC using an empirical bias-corrected ROC curve. The new estimator and its properties are introduced in Section 3. A simulation study is presented in Section 4, followed by an example in Section 5 in which the competing methods are illustrated using data from a study of depression in elderly primary care patients. The paper concludes with a discussion.

2 ESTIMATING AUC USING AN EMPIRICAL BIAS-CORRECTED ROC CURVE

Let T_i denote the continuous test result and let D_i denote the true disease status for the i^{th} subject, $i = 1, 2, \dots, n$, where $D_i = 1$ indicates that the subject has the disease and $D_i = 0$ indicates that subject does not have the disease. Only a subset of the subjects have their disease status verified; let $V_i = 1$ if the i^{th} subject has the true disease status verified, and $V_i = 0$ otherwise. Let X_i be a vector of observed covariates for the i^{th} subject that may be associated with both D_i and V_i . Without loss of generality, suppose that larger values of T are more indicative of disease.

All of the methods reviewed in this section are based on the assumption that verification of disease status is conditionally independent of the true disease status given the test result and the observed covariates. In our notation, $V \perp D | (T, X)$. This is the MAR assumption discussed in the Introduction. The decision to verify the subject's true disease status depends on the true disease status only through X and T .

If all subjects have their disease status verified, i.e., $V_i = 1$, $i = 1, 2, \dots, n$, we have a complete data set. For any cutpoint c , the sensitivity, $\text{Se}(c)$, and specificity, $\text{Sp}(c)$, of the test can be easily estimated by simple proportions as

$$\widehat{Se}_{\text{full}}(c) = \frac{\sum_{i=1}^n I(T_i \geq c) D_i}{\sum_{i=1}^n D_i}, \quad \widehat{Sp}_{\text{full}}(c) = \frac{\sum_{i=1}^n I(T_i < c) (1 - D_i)}{\sum_{i=1}^n (1 - D_i)}. \quad (1)$$

These estimators are unbiased for $\text{Se}(c)$ and $\text{Sp}(c)$, respectively.

If only some subjects are selected to have their disease status verified, the naïve estimators of $\text{Se}(c)$ and $\text{Sp}(c)$ that apply the above methods to only the verified subjects are obtained as

$$\widehat{Se}_{\text{naïve}}(c) = \frac{\sum_{i=1}^n V_i I(T_i \geq c) D_i}{\sum_{i=1}^n V_i D_i}, \quad \widehat{Sp}_{\text{naïve}}(c) = \frac{\sum_{i=1}^n V_i I(T_i < c) (1 - D_i)}{\sum_{i=1}^n V_i (1 - D_i)}. \quad (2)$$

The naïve estimators are unbiased only if the subjects are selected for verification completely at random. Under the less restrictive MAR assumption, the naïve estimators are biased.

Alonzo and Pepe [19] extended Begg and Greenes' method [3] to high-dimensional X , where some components of X may be continuous. The resulting estimators of $\text{Se}(c)$ and $\text{Sp}(c)$ are:

$$\widehat{Se}_{\text{BG}}(c) = \frac{\sum_{i=1}^n I(T_i \geq c) \widehat{\Pr}(D_i = 1 | T_i, X_i)}{\sum_{i=1}^n \widehat{\Pr}(D_i = 1 | T_i, X_i)},$$

$$\widehat{Sp}_{\text{BG}}(c) = \frac{\sum_{i=1}^n I(T_i < c) \widehat{\Pr}(D_i = 0 | T_i, X_i)}{\sum_{i=1}^n \widehat{\Pr}(D_i = 0 | T_i, X_i)}. \quad (3)$$

Parametric models such as logistic regression models can be used to estimate $\Pr(D_i = 1 | T_i, X_i)$ using only data from verified subjects (MAR assumption).

Alonzo and Pepe [19] recognized that a study with verification-biased sampling can be thought of as a study with a two-phase or double-sampling design. They applied methods of prevalence estimation in such studies to the problem of estimating sensitivity and specificity in this setting. The mean score (MS) method proposed by Pepe et al. [23] and Reilly and Pepe [24] (see also [27]) estimates $\Pr(D_i = 1|T_i, X_i)$ by D_i for verified subjects, and by $\widehat{\Pr}(D_i=1|T_i|X_i)$ for unverified subjects. The resulting estimators of $Se(c)$ and $Sp(c)$ are:

$$\begin{aligned} \widehat{Se}_{MS}(c) &= \frac{\sum_{i=1}^n I(T_i \geq c) \{V_i D_i + (1 - V_i) \widehat{\Pr}(D_i = 1|T_i, X_i)\}}{\sum_{i=1}^n \{V_i D_i + (1 - V_i) \widehat{\Pr}(D_i = 1|T_i, X_i)\}}, \\ \widehat{Sp}_{MS}(c) &= \frac{\sum_{i=1}^n I(T_i < c) \{V_i (1 - D_i) + (1 - V_i) \widehat{\Pr}(D_i = 0|T_i, X_i)\}}{\sum_{i=1}^n \{V_i (1 - D_i) + (1 - V_i) \widehat{\Pr}(D_i = 0|T_i, X_i)\}}. \end{aligned} \tag{4}$$

Note that the extended Begg and Greenes estimators estimate $\Pr(D_i = 1|T_i, X_i)$ by $\widehat{\Pr}(D_i=1|T_i|X_i)$ for all subjects.

The inverse probability weighting (IPW) approach weights each verified subject by the inverse of the selection probability to correct verification bias. This method dates back to Horvitz and Thompson [22] and has a long history in the analysis of sample surveys. Let $\pi_i = \Pr(V_i|T_i, X_i)$. For the naïve estimator, if each verified subject is given weight $\widehat{\pi}_i^{-1}$, the inverse of the estimated probability that the subject was selected for verification, the estimators for $Se(c)$ and $Sp(c)$ are

$$\widehat{Se}_{IPW}(c) = \frac{\sum_{i=1}^n V_i I(T_i \geq c) D_i \widehat{\pi}_i^{-1}}{\sum_{i=1}^n V_i D_i \widehat{\pi}_i^{-1}}, \quad \widehat{Sp}_{IPW}(c) = \frac{\sum_{i=1}^n V_i I(T_i < c) (1 - D_i) \widehat{\pi}_i^{-1}}{\sum_{i=1}^n V_i (1 - D_i) \widehat{\pi}_i^{-1}}. \tag{5}$$

The semi-parametric efficient approach (SP) of Alonzo et al. [28] is based on ideas first suggested by Robins et al. [25] and Robins and Rotnitzky [26] and yields estimators that are doubly robust in the sense that they are consistent if either π_i or $\Pr(D|T, X)$ is estimated consistently. The values $Se(c)$ and $Sp(c)$ are estimated as follows:

$$\begin{aligned} \widehat{Se}_{SP}(c) &= \frac{\sum_{i=1}^n I(T_i \geq c) \{V_i D_i + (\widehat{\pi}_i - V_i) \widehat{\Pr}(D_i = 1|T_i, X_i)\} \widehat{\pi}_i^{-1}}{\sum_{i=1}^n \{V_i D_i + (\widehat{\pi}_i - V_i) \widehat{\Pr}(D_i = 1|T_i, X_i)\} \widehat{\pi}_i^{-1}}, \\ \widehat{Sp}_{SP}(c) &= \frac{\sum_{i=1}^n I(T_i < c) \{V_i (1 - D_i) + (\widehat{\pi}_i - V_i) \widehat{\Pr}(D_i = 0|T_i, X_i)\} \widehat{\pi}_i^{-1}}{\sum_{i=1}^n \{V_i (1 - D_i) + (\widehat{\pi}_i - V_i) \widehat{\Pr}(D_i = 0|T_i, X_i)\} \widehat{\pi}_i^{-1}}. \end{aligned} \tag{6}$$

For each of the above methods, when c is varied throughout the real line, an empirical bias-corrected ROC curve is obtained using the pairs $(1 - \widehat{Sp}(c), \widehat{Se}(c))$ [19]. An estimate of the AUC is easily derived empirically. A limitation to this estimator of AUC is that there is no closed-form expression for its variance. Resampling methods such as the bootstrap are needed to make inferences about the AUC.

3 DIRECT ESTIMATION OF AUC

3.1 A brief review of U-statistics

In this subsection we give a very brief account of the U-statistics theory required for our derivations below. See [29] for more details.

Let X be a random variable or vector with distribution F , and let $\theta(F)$ denote a real-valued function defined for F . Suppose that there exists a real-valued function $h(X_1, \dots, X_m)$ such that $E_F(h(X_1, \dots, X_m)) = \theta(F)$ for all F subject only to mild restrictions on h such as continuity and existence of moments. Then for a sample of size $n \geq m$, an unbiased estimator

of $\theta(F)$ can be constructed, namely $U_n = U_n(h) = \frac{\sum_{P_{m,n}} h(X_{i_1}, \dots, X_{i_m})}{n!/(n-m)!}$, where $P_{m,n}$ is the set of all $\frac{n!}{(n-m)!}$ permutations (i_1, i_2, \dots, i_m) of size m chosen from $(1, 2, \dots, n)$. The statistic U_n is called a U-statistic with kernel h . Note that without loss of generality h may be assumed to be a symmetric function of its arguments.

To describe the asymptotic distribution of the U-statistic U_n , we introduce some notation.

Let $\sigma_c^2 = \text{Cov}(h(X_{i_1}, \dots, X_{i_m}), h(X_{j_1}, \dots, X_{j_m}))$, where (i_1, i_2, \dots, i_m) and (j_1, j_2, \dots, j_m) are permutations of size m chosen from $(1, 2, \dots, n)$, and c is the number of integers in $\{i_1, i_2, \dots, i_m\} \cap \{j_1, j_2, \dots, j_m\}$, $0 \leq c \leq m$.

Theorem 1— U_n is an unbiased estimator for $\theta(F)$. If $\sigma_m^2 < \infty$, then

$$\sqrt{n}(U_n - \theta) \xrightarrow{P} N(0, m^2 \sigma_1^2).$$

3.2 Direct estimation of the AUC in the presence of verification bias

In the presence of verification bias, due to the lack of a closed-form expression for the variance, inference for the AUC based on an empirical bias-corrected ROC curve can only be performed with the aid of resampling methods such as the bootstrap. Since these methods are computationally-intensive, it would be convenient to have an estimator similar to the Wilcoxon statistic for which the variance can be derived in closed form. The difficulty of generalizing the Wilcoxon statistic to direct estimation of the AUC in this case lies in the fact that the true disease status of those subjects who have not been administered the gold standard evaluation is not known. Thus it is impossible to divide the whole sample into the two subsamples: diseased and non-diseased. Instead we propose a one sample U-statistic estimator described below.

Let T_i, V_i, D_i and X_i be defined as before. Assume that the $S_i = (T_i, X_i, V_i, D_i)$ are i.i.d. and that $V_i \perp D_i | T_i, X_i, i = 1, \dots, n$ (MAR). So (T_i, X_i, V_i) is observed for everybody, but D_i is observed only if $V_i = 1$. Let λ be the disease prevalence, and let $\pi_i = \Pr(V_i = 1 | T_i, X_i)$ be the verification probability. Also, let $F_1(t)$ be the distribution function of T for subjects with $D_i = 1$, and $F_0(t)$ be the corresponding distribution function for subjects with $D_i = 0$. Finally, let $G_1(t, x)$ be the joint distribution function of T and X for subjects with $D_i = 1$, and $G_0(t, x)$ be the corresponding distribution function for subjects with $D_i = 0$.

Assume for the moment that π_i is known, $i = 1, 2, \dots, n$. The new estimator is motivated by the following observation:

$$\begin{aligned} & E[\pi_i^{-1} \pi_j^{-1} V_i V_j I(T_i > T_j) I(D_i > D_j)] \\ &= E[E[\pi_i^{-1} \pi_j^{-1} V_i V_j I(T_i > T_j) I(D_i > D_j) | T_i, T_j, X_i, X_j]] \\ &= E[I(T_i > T_j) E[I(D_i > D_j) | T_i, T_j, X_i, X_j] E[\pi_i^{-1} V_i | T_i, T_j, X_i, X_j] E[\pi_j^{-1} V_j | T_i, T_j, X_i, X_j]] \\ &= E[I(T_i > T_j) E[I(D_i > D_j) | T_i, T_j, X_i, X_j]] \\ &= E[I(T_i > T_j) I(D_i > D_j)] \\ &= \Pr(D_i = 1) \Pr(D_j = 0) \text{AUC}. \end{aligned}$$

Note that $\pi_i^{-1}\pi_j^{-1}V_iV_jI(T_i>T_j)I(D_i>D_j)$ is computable for every subject even if D_i is unknown. This is because if D_i is unknown, then $V_i = 0$, and therefore the value of the expression is 0 regardless of the value of D_i . So to get an estimator for the AUC, a statistic that has expectation $\Pr(D_i = 1) \Pr(D_j = 0)$ could be used in the denominator. A natural choice is $\pi_i^{-1}\pi_j^{-1}V_iV_jI(D_i>D_j)$, since $E[\pi_i^{-1}\pi_j^{-1}V_iV_jI(D_i>D_j)] = \Pr(D_i=1)\Pr(D_j=0)$ using an argument similar to that given above.

We propose the following estimator of the AUC in the presence of verification bias:

$$\begin{aligned} \widehat{AUC} &= \frac{\sum_{i=1}^n \sum_{j=1}^n \pi_i^{-1} \pi_j^{-1} V_i V_j I(T_i > T_j) I(D_i > D_j)}{\sum_{i=1}^n \sum_{j=1}^n \pi_i^{-1} \pi_j^{-1} V_i V_j I(D_i > D_j)} \\ &= \frac{\sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} \pi_i^{-1} \pi_j^{-1} V_i V_j [I(T_i > T_j) I(D_i > D_j) + I(T_i < T_j) I(D_i < D_j)]}{\sum_{i=1}^n \sum_{j=1}^n \frac{1}{2} \pi_i^{-1} \pi_j^{-1} V_i V_j [I(D_i > D_j) + I(D_i < D_j)]}. \end{aligned} \tag{7}$$

Note that the reason for writing the estimator in the second form (symmetric form) is to express it as a function of U-statistics.

The new estimator uses inverse probability weighting to correct verification bias, where the weight $\pi_i^{-1}\pi_j^{-1}$ is attached to all possible pairs of verified subjects. Similar to IPW estimators, the new estimator \widehat{AUC} is not unbiased, but it is consistent.

Theorem 2—The new estimator \widehat{AUC} is consistent.

The asymptotic distribution of \widehat{AUC} is given in the following theorem:

Theorem 3—Let $A = \left(\frac{1}{\lambda(1-\lambda)}, -\frac{AUC}{\lambda(1-\lambda)} \right)$. Then $\sqrt{n}(\widehat{AUC} - AUC) \xrightarrow{d} N(0, \sigma^2)$, where $\sigma^2 = A^T \Sigma A$ and

$$\Sigma = \begin{pmatrix} \text{Cov}(\varphi(S_i, S_j), \varphi(S_i, S_k)) & \text{Cov}(\varphi(S_i, S_j), \psi(S_i, S_k)) \\ \text{Cov}(\varphi(S_i, S_j), \psi(S_i, S_k)) & \text{Cov}(\psi(S_i, S_j), \psi(S_i, S_k)) \end{pmatrix}$$

can be expressed in terms of λ, F_0, F_1, G_0 and G_1 .

The Appendix contains proofs of Theorems 2 and 3 and the derivations of the covariance terms in Theorem 3.

Based on this theorem, one can estimate the variance of \widehat{AUC} by substituting $\hat{\lambda}$, which is estimated by $\sum_{i=1}^n V_i D_i \pi_i^{-1} / \sum_{i=1}^n V_i \pi_i^{-1}$, and the EDFs $\hat{F}_0, \hat{F}_1, \hat{G}_0$ and \hat{G}_1 for their respective population quantities and replacing integrals with sums.

The derivation of the asymptotic properties of \widehat{AUC} assumed that the $\pi_i = \Pr(V_i = 1 | T_i, X_i)$, $i = 1, 2, \dots, n$, were known. It is not unusual in practice to encounter this situation [19]. Otherwise, $\hat{\pi}_i$ (obtained by, say, logistic regression) can be substituted for π_i in the expressions for the estimator and its variance. Although the resulting estimated variance does not properly account for the variation in the $\hat{\pi}_i$, this should not be a significant problem in large samples since the variance in the $\hat{\pi}_i$ will be relatively small. This is illustrated in the simulation study presented in Section 4.

The new estimator has an interesting property: it is identical to the empirical AUC based on the IPW approach.

Theorem 4—The estimator \widehat{AUC} is equivalent to the empirical AUC based on the empirical bias-corrected ROC curve obtained using the IPW approach.

The proof of Theorem 4 is also given in the Appendix.

4 SIMULATION STUDIES

4.1 Study design

In this section, the finite-sample behavior of the new estimator of the AUC relative to those of existing estimators based on the empirical bias-corrected ROC curve is investigated via simulation. The existing estimators include the naïve estimator and those based on methods for constructing an empirical bias-corrected ROC curve: modified Begg and Greenes (BG), mean score (MS), and the semi-parametric efficient approach (SP). Note that the new direct estimator is equivalent to the area under the empirical bias-corrected ROC curve derived using inverse probability weighting (IPW).

The simulation set-up is similar to that of Alonzo et al. [19,28]. The disease is considered to arise from two underlying continuous disease processes, which remain subclinical until some function of the processes exceeds a certain threshold, at which point the disease becomes apparent. In particular, two independent random variables $Z_1 \sim N(0, 0.5)$ and $Z_2 \sim N(0, 0.5)$ were generated, and the disease indicator D was specified as $D = I[g(Z_1, Z_2) > r]$. Thus, by varying $g(Z_1, Z_2)$ one can consider different disease processes, and by varying r one can consider different disease prevalences. The continuous diagnostic test result T was assumed to be related to D through Z_1 and Z_2 : $T = \alpha_1 Z_1 + \beta_1 Z_2 + \varepsilon_1$, where $\varepsilon_1 \sim N(0, 0.25)$ and is independent of Z_1 and Z_2 . A single covariate X was chosen to be related to the two separate components of the disease process: $X = \alpha_2 Z_1 + \beta_2 Z_2 + \varepsilon_2$, where $\varepsilon_2 \sim N(0, 0.25)$ and is also independent of Z_1 and Z_2 . By varying α_1 , α_2 , β_1 , and β_2 , one can vary the extent to which the test result and the covariates capture the different components of the underlying disease process, as well as the correlations between the test result and the covariates. The values also affect the discriminatory abilities of T and X with respect to D . Finally, the verification probability $h(T, X)$ was chosen to be a specified function of T and X in keeping with the MAR assumption.

Using this simulation set-up, we verified the results of Alonzo and Pepe [19] who showed that, when the models for verification and disease are correctly specified, the BG, MS, SP, and IPW (new) methods have minimal bias, with the BG and MS estimators being typically more precise than the SP and IPW (new) estimators. When the disease model is misspecified, however, the BG and MS estimators exhibit substantial bias, whereas the SP and IPW (new) estimators have very little bias and are similar with respect to variance.

4.2 Performance of the variance estimator

Let $g(Z_1, Z_2) = Z_1 + Z_2$ and $h(T, X) = \delta + (1 - \delta)I(T > t^q)$, where $0 \leq \delta < 1$ and t^q is the q^{th} quantile of the distribution of T . The decision to verify the disease status of the subject thus does not depend on X . In this simulation study, different values of α_1 and β_1 are considered for $T = \alpha_1 Z_1 + \beta_1 Z_2 + \varepsilon_1$ and $\alpha_2 = \beta_2 = 1$. The value of δ is chosen to be 0.20 and q is chosen to be 80. Therefore, all subjects with a value of T above the 80th percentile, and 20% of all other subjects, are selected to have their disease status verified. The probability of verification in the population is thus 0.36. In the estimation procedure, the correct models were assumed to hold for verification ($V|T$) and disease ($D|T, X$). For disease, a generalized linear model for D given T and X with probit link is the correct model [19].

The performance of the asymptotic variance estimator for the AUC as a function of sample size, disease prevalence, and the value of the AUC was investigated. The threshold value r was varied to yield different disease prevalences. Also, by varying α_1 and β_1 , the extent to which the test result T captures the different components of the underlying disease process is varied, yielding different values of the true AUC. The simulation variance of the AUC was calculated from 5000 realizations and can be considered to reflect the true variability of the estimator. The performance of the variance estimator was assessed by examining the ratio of the estimate (averaged over the 5000 realizations) to the simulation variance. To assess the impact of using the estimated verification probabilities in place of the true verification probabilities in the variance formula, we present the results for both cases. For comparison, we also estimated the variances using the bootstrap approach with 500 bootstrap samples. Table 1 provides the variance ratios for a range of sample sizes and AUC values when the disease prevalence is 0.3 or 0.5. The asymptotic variance formula performs very well for sample sizes as small as 100 when the disease prevalence is 0.5, and for sample sizes as small as 200 when the disease prevalence is 0.3. There is virtually no difference in the performance of the formula when the true and the estimated verification probabilities are used, which indicates that the variation due to the estimation of the verification probabilities is very small relative to the variation of the AUC even for sample sizes as small as 100. When the sample sizes are relatively small, the asymptotic variance formula tends to underestimate the variance, while the bootstrap approach tends to overestimate the variance.

5 Study of Depression in Elderly Primary Care Patients

We illustrate our proposed methodology using data from a longitudinal study of depression in elderly patients (age ≥ 65) recruited from primary care practices in Monroe County, New York. At the intake evaluation, 708 patients underwent a comprehensive diagnostic assessment for depression using the Structured Clinical Interview for DSM-IV (SCID), an intensive examiner-based assessment that can be considered as a practical gold standard for this purpose [30]. Depression was defined based on the SCID as major or minor depression, actively symptomatic (i.e., either current or partially remitted); 249 patients were classified as having depression and 459 patients were classified as not having depression. Other information collected as part of this study included the Hamilton Depression Rating Scale (HAM-D), a 24-item observer-rated scale designed to measure the severity of depressive symptoms [31]. In this example, the utility of the HAM-D as a screening marker for the diagnosis of depression will be evaluated. The HAM-D takes approximately 15–20 minutes to administer, compared to 1–3 hours for the SCID.

Data for both the SCID and the HAM-D were collected from all participating patients in this study; therefore, we used a subset of these data that resemble data that would be obtained from a two-phase design. In this subset, HAM-D results are available for all patients, but SCID diagnoses are available only for certain patients selected according to the following mechanism:

$$\Pr(\text{SCID available})=0.05+0.50I[\text{HAM} - \text{D}>7]+0.45I[\text{CIRS}>7]I[\text{Age}<75],$$

where the CIRS is the total score on the Cumulative Illness Rating Scale, a reliable and valid measure of medical burden that quantifies the amount of pathology in each organ system [32]. Thus, the verification mechanism preferentially selected patients who had a HAM-D score > 7 or patients under the age of 75 with a relatively high cumulative illness burden. Using this mechanism, 289 of the 708 patients (41%) were selected for SCID verification of the depression diagnosis.

We consider estimation of the AUC of the HAM-D for screening for depression and treat age, gender, years of education, and CIRS total score as covariates (i.e., $D = \text{SCID}$ diagnosis, $T = \text{HAM-D}$ and $X = [\text{age, gender, years of education, CIRS total score}]$ in terms of previous notation). The AUC was estimated using the new direct estimator, the naïve estimator, and estimators derived from the empirical bias-corrected ROC curves, including BG, MS, and SP; recall that the new direct estimator is equivalent to the IPW estimator. Since the full data are available (in addition to the selected subset), the estimators in the setting of verification bias can be compared to the “full data” estimator, which is not subject to this bias.

The BG, MS and SP estimators require a model for $\Pr(D|T, X)$. A logistic regression model was used for this purpose assuming linear relationships between log-odds of depression and age, years of education, and CIRS total score. The IPW (new) and SP estimators require a model for $\Pr(V|T, X)$, hence we used the observed fractions of subjects falling in the 8 categories defined by the combinations of $I[\text{HAM-D} > 7]$, $I[\text{CIRS} > 7]$ and $I[\text{Age} < 75]$. The resulting estimates are presented in Table 2. The confidence intervals for the AUC were computed using 500 replications of bootstrap resampling as well as the asymptotic variance result from Theorem 3 for the IPW (new) estimate. As expected, the naïve estimate of the AUC is noticeably biased. Although the true model for $\Pr(D|T, X)$ in this case is not known, it is likely that there is some degree of model misspecification present. The BG and MS estimators are biased. In this example, the verification mechanism is well understood and the IPW (new) approach yields an estimate that is quite close to the full data estimate. Surprisingly, the SP approach yields an estimate that differs somewhat from the full data estimate in this example. Also, for the new estimate, the confidence intervals obtained via the bootstrap and Theorem 3 are virtually identical.

6 DISCUSSION

In this paper we propose a direct estimator of AUC in the presence of verification bias when the test result is continuous. The estimator is based on inverse probability weighting and has a simple closed-form expression. Because the estimator is a function of U-statistics, a closed-form expression for its asymptotic variance could also be derived.

Several methods exist for direct estimation of the AUC, but these have been derived for the case when the test result T is ordinal and the relevant covariates X are categorical. Recent work by Alonzo and Pepe [19] has developed methods for estimating the AUC for the case where the test result is continuous and the covariates can be continuous. The general approach is to first derive the empirical bias-corrected ROC curves, and then compute the area under these curves using the trapezoidal rule. A limitation of this approach is that there is no closed-form expression for the variance of the estimator and resampling techniques are needed for inference. Our proposed direct estimator of the AUC is easily computed, as is its approximate variance. The variance formula appeared to work quite well in our simulation studies, comparing the results from the formula with the simulation variance results. An interesting result is that the new estimator is equivalent to the AUC computed using Alonzo and Pepe’s IPW approach [19]. It should be noted that the variance formula for the new estimator assumes that the observations are independent. For situations where this assumption is violated, resampling techniques can be used to accommodate the dependence. Alonzo and Pepe [19] provide an example from a study of neonatal hearing screening in which observations on the test result (T) and diagnosis (D) were obtained from both ears in most of the subjects, resulting in dependent or clustered data. Our variance formula cannot be applied in this situation.

The results of our simulation studies show that when the models for disease and verification are correctly specified, all existing methods for bias correction perform quite well. The BG and MS methods appear to be somewhat more efficient than the SP and IPW (new) methods, as noted by Alonzo and Pepe [19]. On the other hand, the SP method has the advantage of being doubly robust to model misspecification. Also, the IPW (new) method requires correct specification only for the model for verification of true disease status. It is often the case in practice that the verification mechanism is well understood or can be controlled by the investigators, in which case model misspecification is less of an issue. For the BG and MS methods, the model for disease must be correctly specified, which may be more challenging in practice. When this model is incorrectly specified, the estimators can have significant bias.

Rotnitzky et al. [20] recently proposed a direct estimator of the AUC that is doubly robust and can be applied in cases where the mechanism for missing true disease status is nonignorable. Under the MAR assumption, this estimator is essentially the same as the estimator based on the SP method. Indeed, both approaches are based on the idea of replacing the disease status D_i with

$$\widehat{D}_i = \widehat{Pr}(D_i=1|T_i, X_i) + (V_i/\pi_i)(D_i - \widehat{Pr}(D_i=1|T_i, X_i))$$

for all subjects. In fact, using the same strategy as that in the proof of Theorem 4, the empirical AUC based on the SP approach can be derived as

$$\widehat{AUC}_{SP} = \frac{\sum_i \sum_j (1 - \widehat{D}_i) \widehat{D}_j I_{i,j}}{\sum_i \sum_j (1 - \widehat{D}_i) \widehat{D}_j}$$

where $I_{i,j} = I(T_i < T_j) + \frac{1}{2} I(T_i = T_j)$. This is exactly the same as the estimator of Rotnitzky et al. [20], except that the latter estimator excludes the terms in the above expression for which $i = j$. The numerical difference between the two estimators in practice is very small and they have the same asymptotic properties. Hence, this is another method with a closed-form expression for the estimator and its asymptotic variance. We (and others [19]) have shown that these two estimators (SP and IPW) behave quite similarly when the model for verification is correctly specified, regardless of whether the model for disease is correctly specified. The SP estimator will be superior when the model for disease is correctly specified but the model for verification is incorrectly specified. If, however, the verification mechanism is well understood, the IPW method does not require the extra step of specifying (and fitting) a model for disease. This step may not be straightforward in some situations (e.g., when the disease prevalence is low and the resulting sparse data creates problems in model-fitting).

Acknowledgments

Dr. Michael P. McDermott acknowledges support from the University of Rochester CTSI, Grant Number 1 UL1 RR024160-01 from the National Center for Research Resources, a component of the National Institutes of Health (NIH) and the NIH Roadmap for Medical Research. Dr. Jeffrey M. Lyness acknowledges support from the National Institute of Mental Health (NIMH), Grant Numbers 1 R01 MH061429 and K24 MH071509.

References

1. Bamber D. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975; 12:387–415.
2. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*. 1982; 143:29–36. [PubMed: 7063747]
3. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. *Biometrics*. 1983; 39:207–215. [PubMed: 6871349]
4. Begg CB. Biases in the assessment of diagnostic tests. *Statistics in Medicine*. 1987; 6:411–423. [PubMed: 3114858]
5. Zhou XH. Effect of verification bias on positive and negative predictive values. *Statistics in Medicine*. 1994; 13:1737–1745. [PubMed: 7997707]
6. Zhou XH. Correcting for verification bias in studies of a diagnostic test's accuracy. *Statistical Methods in Medical Research*. 1998; 7:337–353. [PubMed: 9871951]
7. Ransohoff DF, Feinstein AR. Problems of spectrum and bias in evaluating the efficacy of diagnostic tests. *New England Journal of Medicine*. 1978; 299:926–930. [PubMed: 692598]
8. Gray R, Begg CB, Greenes RA. Construction of receiver operating characteristic curves when disease verification is subject to selection bias. *Medical Decision Making*. 1984; 4:151–164. [PubMed: 6472063]
9. Hunink MGM, Richardson DK, Doubilet PM, Begg CB. Testing for fetal pulmonary maturity: ROC analysis involving covariates, verification bias, and combination testing. *Medical Decision Making*. 1990; 10:201–211. [PubMed: 2370827]
10. Zhou XH. Maximum likelihood estimators of sensitivity and specificity corrected for verification bias. *Communications in Statistics: Theory and Methods*. 1993; 22:3177–3198.
11. 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]
12. Baker SG. Evaluating multiple diagnostic tests with partial verification. *Biometrics*. 1995; 51:330–337. [PubMed: 7539300]
13. Zhou XH, Rodenberg C. Estimating the ROC curve in the presence of non-ignorable verification bias. *Communications in Statistics: Theory and Methods*. 1998; 27:635–657.
14. Toledano AY, Gatsonis C. Generalized estimating equations for ordinal categorical data: arbitrary patterns of missing responses and missingness in a key covariate. *Biometrics*. 1999; 55:488–496. [PubMed: 11318205]
15. Rodenberg C, Zhou XH. ROC curve estimation when covariates affect the verification process. *Biometrics*. 2000; 56:1256–1262. [PubMed: 11129488]
16. Kosinski AS, Barnhart HX. Accounting for nonignorable verification bias in assessment of diagnostic tests. *Biometrics*. 2003; 59:163–171. [PubMed: 12762453]
17. 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.
18. Little, RJA.; Rubin, DB. *Statistical Analysis with Missing Data*. 2. Wiley; Hoboken: 2002.
19. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society, Series C*. 2005; 54:173–190.
20. 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.
21. Carroll, RJ.; Ruppert, D.; Stefanski, LA. *Measurement Error in Nonlinear Models*. Chapman & Hall; London: 1995.
22. Horvitz DG, Thompson DJ. A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*. 1952; 47:663–685.
23. Pepe MS, Reilly M, Fleming TR. Auxiliary outcome data and the mean score method. *Journal of Statistical Planning and Inference*. 1994; 42:137–160.

24. Reilly M, Pepe MS. A mean score method for missing and auxiliary covariate data in regression models. *Biometrika*. 1995; 82:299–314.
25. Robins JM, Rotnitzky A, Zhao LP. Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*. 1994; 89:846–866.
26. Robins JM, Rotnitzky A. Semiparametric efficiency in multivariate regression models with missing data. *Journal of the American Statistical Association*. 1995; 90:122–129.
27. Clayton D, Spiegelhalter D, Dunn G, Pickles A. Analysis of longitudinal binary data from multi-phase sampling. *Journal of the Royal Statistical Society, Series B*. 1998; 60:71–87.
28. Alonzo TA, Pepe MS, Lumley T. Estimating disease prevalence in two-phase studies. *Biostatistics*. 2003; 4:313–326. [PubMed: 12925524]
29. Lee, AJ. *U-Statistics: Theory and Practice*. Marcel Dekker; New York: 1990.
30. Spitzer, RL.; Gibbon, M.; Williams, JBW. *Structured Clinical Interview for Axis I DSM-IV Disorders*. Biometrics Research Department, New York State Psychiatric Institute; 1994.
31. Williams JBW. A structured interview guide for the Hamilton Depression Rating Scale. *Archives of General Psychiatry*. 1988; 45:742–747. [PubMed: 3395203]
32. Linn BS, Linn MW, Gurel L. Cumulative illness rating scale. *Journal of the American Geriatrics Society*. 1968; 16:622–626. [PubMed: 5646906]

APPENDIX

Proof of Theorem 2

Let $\varphi(S_i, S_j) = \frac{1}{2} \frac{1}{\pi_i \pi_j} V_i V_j [I(T_i > T_j)I(D_i > D_j) + I(T_i < T_j)I(D_i < D_j)]$, and let $\psi(S_i, S_j) = \frac{1}{2} \frac{1}{\pi_i \pi_j} V_i V_j [I(D_i > D_j) + I(D_i < D_j)]$. As computed in Section 3.2, we have $E[\varphi(S_i, S_j)] = \lambda(1 - \lambda)AUC$, where $\lambda = \Pr(D_i = 1)$.

By Theorem 1, we have $\frac{\sum_{i \neq j} \varphi(S_i, S_j)}{n(n-1)} \xrightarrow{P} \lambda(1 - \lambda)AUC$. Also, since

$$E[\psi(S_i, S_j)] = \lambda(1 - \lambda), \quad \frac{\sum_{i \neq j} \psi(S_i, S_j)}{n(n-1)} \xrightarrow{P} \lambda(1 - \lambda). \quad \text{Hence } \widehat{AUC} = \frac{\sum_{i \neq j} \varphi(S_i, S_j)}{\sum_{i \neq j} \psi(S_i, S_j)} \xrightarrow{P} AUC.$$

Proof of Theorem 3

Let $U = \sum_{i \neq j} \varphi(S_i, S_j)$ and $V = \sum_{i \neq j} \psi(S_i, S_j)$. By Theorem 1, with $m = 2$, $\sqrt{n}((U, V)^T - \mu) \xrightarrow{d} N(0, 4 \sum)$, where $\mu = (\lambda(1 - \lambda)AUC, \lambda(1 - \lambda))^T$. Therefore, by the multivariate δ -method, $\sqrt{n}(\widehat{AUC} - AUC) \xrightarrow{d} N(0, \sigma^2)$.

The matrix S can be expressed in terms of λ , F_0 , F_1 , G_0 , and G_1 . Since

$$\begin{aligned} \text{Cov}(\varphi(S_i, S_j), \varphi(S_i, S_k)) &= E[\varphi(S_i, S_j)\varphi(S_i, S_k)] - \lambda^2(1 - \lambda)^2 AUC^2, \\ \text{Cov}(\psi(S_i, S_j), \psi(S_i, S_k)) &= E[\psi(S_i, S_j)\psi(S_i, S_k)] - \lambda^2(1 - \lambda)^2, \\ \text{Cov}(\varphi(S_i, S_j), \psi(S_i, S_k)) &= E[\varphi(S_i, S_j)\psi(S_i, S_k)] - \lambda^2(1 - \lambda)^2 AUC, \end{aligned}$$

in order to compute Σ , we need to compute

$$E[\varphi(S_i, S_j)\varphi(S_i, S_k)], E[\psi(S_i, S_j)\psi(S_i, S_k)], \text{ and } E[\varphi(S_i, S_j)\psi(S_i, S_k)].$$

Computation of $E[\varphi(S_i, S_j)\varphi(S_i, S_k)]$

$$\begin{aligned}
 & 4E[\varphi(S_i, S_j)\varphi(S_i, S_k)] \\
 & = E[\pi_i^{-1}[I(T_i > T_j)I(D_i > D_j)I(T_i > T_k)I(D_i > D_k) \\
 & \quad + I(T_i < T_j)I(D_i < D_j)I(T_i < T_k)I(D_i < D_k)]] \\
 & = \lambda(1 - \lambda)^2 E[\pi_i^{-1}I(T_i > T_j)I(T_i > T_k)|D_i=1, D_j=D_k=0] \\
 & \quad + \lambda^2(1 - \lambda)E[\pi_i^{-1}I(T_i < T_j)I(T_i < T_k)|D_i=0, D_j=D_k=1] \\
 & = \lambda(1 - \lambda)^2 E[\pi_i^{-1} \int I(T_i > T_j)dF_0(T_j) \int I(T_i > T_k)dF_0(T_k)|D_i=1] \\
 & \quad + \lambda^2(1 - \lambda)E[\pi_i^{-1} \int I(T_i < T_j)dF_1(T_j) \int I(T_i < T_k)dF_1(T_k)|D_i=0] \\
 & = \lambda(1 - \lambda)^2 \int \pi_i^{-1} F_0^2(T_i) dG_1(T_i, X_i) + \lambda^2(1 - \lambda) \int \pi_i^{-1} (1 - F_1(T_i))^2 dG_0(T_i, X_i)
 \end{aligned}$$

Hence

$$Cov(\varphi(S_i, S_j), \varphi(S_i, S_k)) = \frac{1}{4}\lambda(1 - \lambda)^2 \int \pi_i^{-1} F_0^2(T_i) dG_1(T_i, X_i) + \frac{1}{4}\lambda^2(1 - \lambda) \int \pi_i^{-1} (1 - F_1(T_i))^2 dG_0(T_i, X_i) - \lambda^2(1 - \lambda)^2 AUC^2$$

Similarly,

$$Cov(\psi(S_i, S_j), \psi(S_i, S_k)) = \frac{1}{4}\lambda(1 - \lambda)^2 \int \pi_i^{-1} dG_1(T_i, X_i) + \frac{1}{4}\lambda^2(1 - \lambda) \int \pi_i^{-1} dG_0(T_i, X_i) - \lambda^2(1 - \lambda)^2,$$

and

$$Cov(\varphi(S_i, S_j), \psi(S_i, S_k)) = \frac{1}{4}\lambda(1 - \lambda)^2 \int \pi_i^{-1} F_0(T_i) dG_1(T_i, X_i) + \frac{1}{4}\lambda^2(1 - \lambda) \int \pi_i^{-1} (1 - F_1(T_i)) dG_0(T_i, X_i) - \lambda^2(1 - \lambda)^2 AUC$$

Thus,

$$\begin{aligned}
 Cov(\varphi(S_i, S_j), \varphi(S_i, S_k)) & = \frac{1}{4}\lambda(1 - \lambda)^2 \int \pi_i^{-1} F_0^2(T_i) dG_1(T_i, X_i) \\
 & \quad + \frac{1}{4}\lambda^2(1 - \lambda) \int \pi_i^{-1} (1 - F_1(T_i))^2 dG_0(T_i, X_i) - \lambda^2(1 - \lambda)^2 AUC^2, \\
 Cov(\psi(S_i, S_j), \psi(S_i, S_k)) & = \frac{1}{4}\lambda(1 - \lambda)^2 \int \pi_i^{-1} dG_1(T_i, X_i) \\
 & \quad + \frac{1}{4}\lambda^2(1 - \lambda) \int \pi_i^{-1} dG_0(T_i, X_i) - \lambda^2(1 - \lambda)^2, \text{ and} \\
 Cov(\varphi(S_i, S_j), \psi(S_i, S_k)) & = \frac{1}{4}\lambda(1 - \lambda)^2 \int \pi_i^{-1} F_0(T_i) dG_1(T_i, X_i) \\
 & \quad + \frac{1}{4}\lambda^2(1 - \lambda) \int \pi_i^{-1} (1 - F_1(T_i)) dG_0(T_i, X_i) - \lambda^2(1 - \lambda)^2 AUC.
 \end{aligned}$$

Proof of Theorem 4

Let $c_1 < \dots < c_n$ be the ordered test results t_1, \dots, t_n . Then the points forming the empirical ROC curve obtained using the IPW approach are

$$\left(1 - \frac{\sum V_i I(T_i < c_k) (1 - D_i) \widehat{\pi}_i^{-1}}{\sum V_i (1 - D_i) \widehat{\pi}_i^{-1}}, \frac{\sum V_i I(T_i \geq c_k) D_i \widehat{\pi}_i^{-1}}{\sum V_i D_i \widehat{\pi}_i^{-1}} \right), \quad k=1, 2, \dots, n.$$

Thus the area under the empirical ROC curve is

$$\begin{aligned} & \sum_{k=1}^{n-1} \frac{1}{2} \left(\frac{\sum V_i I(T_i < c_{k+1})(1-D_i) \widehat{\pi}_i^{-1}}{\sum V_i (1-D_i) \widehat{\pi}_i^{-1}} - \frac{\sum V_i I(T_i < c_k)(1-D_i) \widehat{\pi}_i^{-1}}{\sum V_i (1-D_i) \widehat{\pi}_i^{-1}} \right) \\ & \cdot \left(\frac{\sum V_i I(T_i \geq c_k) D_i \widehat{\pi}_i^{-1}}{\sum V_i D_i \widehat{\pi}_i^{-1}} + \frac{\sum V_i I(T_i \geq c_{k+1}) D_i \widehat{\pi}_i^{-1}}{\sum V_i D_i \widehat{\pi}_i^{-1}} \right) \\ & = \frac{\frac{1}{2} \sum_{k=1}^{n-1} (\sum V_i [I(T_i < c_{k+1}) - I(T_i < c_k)] (1-D_i) \widehat{\pi}_i^{-1})}{\sum V_i (1-D_i) \widehat{\pi}_i^{-1} \sum V_i D_i \widehat{\pi}_i^{-1}} \\ & + \frac{\frac{1}{2} \sum_{k=1}^{n-1} (\sum V_i [I(T_i \geq c_k) + I(T_i \geq c_{k+1})] D_i \widehat{\pi}_i^{-1})}{\sum V_i (1-D_i) \widehat{\pi}_i^{-1} \sum V_i D_i \widehat{\pi}_i^{-1}}. \end{aligned}$$

Since $\sum_i V_i (1 - D_i) \widehat{\pi}_i^{-1} \sum_i V_i D_i \widehat{\pi}_i^{-1} = \sum_i \sum_j \widehat{\pi}_i^{-1} \widehat{\pi}_j^{-1} V_i V_j I(D_j > D_i)$, the denominator of the above expression equals the denominator of our estimator. It is therefore enough to show that the numerators are the same. Now

$$\begin{aligned} & \sum_{k=1}^{n-1} \frac{1}{2} (\sum_i V_i [I(T_i < c_{k+1}) - I(T_i < c_k)] (1 - D_i) \widehat{\pi}_i^{-1}) \\ & \cdot (\sum_j V_j [I(T_j \geq c_k) + I(T_j \geq c_{k+1})] D_j \widehat{\pi}_j^{-1}) \\ & = \frac{1}{2} \sum_i \sum_{j=k=1}^{n-1} V_i [I(T_i < c_{k+1}) - I(T_i < c_k)] (1 - D_i) \widehat{\pi}_i^{-1} \\ & \cdot V_j [I(T_j \geq c_k) + I(T_j \geq c_{k+1})] D_j \widehat{\pi}_j^{-1}. \end{aligned}$$

Note that $I(T_i < c_{k+1}) - I(T_i < c_k) = 0$ if $T_i < c_k$ or $T_i > c_{k+1}$, and is equal to 1 if $T_i = c_k$. Furthermore, if $T_i = c_k$, then since $T_i \neq T_j$ for $j \neq i$, $T_j \geq c_k = T_i$ is equivalent to $T_j \geq c_{k+1}$, hence

$$\sum_{k=1}^{n-1} V_i [I(T_i < c_{k+1}) - I(T_i < c_k)] (1 - D_i) \widehat{\pi}_i^{-1} V_j [I(T_j \geq c_k) + I(T_j \geq c_{k+1})] D_j \widehat{\pi}_j^{-1} = 2 V_i (1 - D_i) \widehat{\pi}_i^{-1} V_j D_j \widehat{\pi}_j^{-1} I(T_j > T_i)$$

The numerator is thus equal to

$$\frac{1}{2} \sum_i \sum_j 2 V_i (1 - D_i) \widehat{\pi}_i^{-1} V_j D_j \widehat{\pi}_j^{-1} I(T_j > T_i) = \sum_i \sum_j \widehat{\pi}_i^{-1} \widehat{\pi}_j^{-1} V_i V_j I(D_j > D_i) I(T_j > T_i),$$

which is exactly the numerator of our estimator.

Table 1

The performance of the asymptotic variance formula using the estimated verification probabilities, the asymptotic variance formula using the true verification probabilities and the bootstrap variance, expressed as the ratio of the estimated variance from the formula to the simulation variance (5000 realizations).

	Disease Prevalence	Sample Size	True AUC							
			0.50	0.63	0.71	0.77	0.83	0.94	0.98	
Estimated Verification Probabilities	0.3	100	0.81	0.84	0.81	0.87	0.85	0.94	1.01	
		200	0.97	0.91	0.95	0.92	0.97	0.96	1.03	
		500	1.00	0.99	0.96	0.99	0.98	1.01	1.03	
		1000	1.00	0.99	1.00	1.01	0.99	1.00	1.02	
		5000	0.99	1.00	0.96	0.99	1.02	1.01	0.99	
	0.5	100	0.96	0.94	0.94	0.88	0.99	1.01	1.04	
		200	1.00	0.98	0.95	1.00	1.06	1.06	1.04	
		500	1.00	0.98	1.03	0.94	0.96	1.03	1.03	
		1000	1.02	1.02	0.99	1.02	1.00	1.02	0.97	
		5000	0.99	0.98	1.00	1.00	1.03	1.01	1.00	
	True Verification Probabilities	0.3	100	0.81	0.85	0.81	0.87	0.86	0.94	1.00
			200	0.96	0.91	0.94	0.91	0.96	0.96	1.03
			500	0.99	0.98	0.96	0.98	0.98	1.00	1.02
			1000	1.00	0.99	1.01	1.01	1.00	1.00	1.02
			5000	1.00	1.00	0.99	0.99	1.02	1.01	0.99
0.5		100	0.94	0.92	0.93	0.87	0.97	1.00	1.01	
		200	0.99	0.98	0.93	0.98	1.05	1.04	1.02	
		500	0.99	0.98	1.03	0.92	0.95	1.02	1.02	
		1000	1.01	1.02	1.01	1.01	1.00	1.01	0.97	
		5000	0.99	0.98	1.00	0.97	1.03	1.03	1.00	
Bootstrap		0.3	100	1.11	1.06	0.99	1.04	1.00	1.11	1.21
			200	1.09	1.02	1.04	1.00	1.04	1.03	1.14
			500	1.04	1.03	1.00	1.01	1.01	1.03	1.08
			1000	1.00	0.98	1.00	1.00	0.99	1.00	1.01
			5000	1.00	0.98	1.00	1.00	0.99	1.00	1.01

Disease Prevalence	Sample Size	True AUC							
		0.50	0.63	0.71	0.77	0.83	0.94	0.98	
	5000	0.99	1.00	1.00	1.00	1.01	1.00	1.00	1.01
0.5	100	1.20	1.17	1.15	1.10	1.24	1.27	1.27	1.34
	200	1.08	1.07	1.03	1.08	1.15	1.17	1.17	1.21
	500	1.02	1.00	1.05	0.96	0.97	1.06	1.06	1.10
	1000	1.01	0.99	0.98	1.01	1.13	0.99	1.05	1.05
	5000	1.00	1.00	1.00	1.00	1.02	1.00	1.00	1.00

Table 2

Estimates of the AUC using data from a study of depression in elderly primary care patients.

Method	Estimated AUC	Bootstrap 95% Confidence Interval
Full Data	0.84	(0.81, 0.87)
Naïve	0.79	(0.74, 0.84)
BG	0.79	(0.73, 0.85)
MS	0.80	(0.74, 0.86)
SP	0.89	(0.83, 0.95)
IPW (new)	0.84	(0.77, 0.91)
		* (0.77, 0.91)

* Note: Confidence interval based on the asymptotic distribution (Theorem 3).