Gianfranco Adimari* and Monica Chiogna

Nearest-Neighbor Estimation for ROC Analysis under Verification Bias

Abstract: For a continuous-scale diagnostic test, the receiver operating characteristic (ROC) curve is a popular tool for displaying the ability of the test to discriminate between healthy and diseased subjects. In some studies, verification of the true disease status is performed only for a subset of subjects, possibly depending on the test result and other characteristics of the subjects. Estimators of the ROC curve based only on this subset of subjects are typically biased; this is known as verification bias. Methods have been proposed to correct verification bias, in particular under the assumption that the true disease status, if missing, is missing at random (MAR). MAR assumption means that the probability of missingness depends on the true disease status only through the test result and observed covariate information. However, the existing methods require parametric models for the (conditional) probability of disease and/or the (conditional) probability of verification, and hence are subject to model misspecification: a wrong specification of such parametric models can affect the behavior of the estimators, which can be inconsistent. To avoid misspecification problems, in this paper we propose a fully nonparametric method for the estimation of the ROC curve of a continuous test under verification bias. The method is based on nearest-neighbor imputation and adopts generic smooth regression models for both the probability that a subject is diseased and the probability that it is verified. Simulation experiments and an illustrative example show the usefulness of the new method. Variance estimation is also discussed.

Keywords: diagnostic tests; missing data imputation; sensitivity; specificity

DOI 10.1515/ijb-2014-0014

1 Introduction

The evaluation of the ability of a diagnostic or a screening test to separate diseased from non-diseased subjects is a crucial issue in modern medicine. In fact, before applying a test in a clinical setting, rigorous statistical assessment of its performance in discriminating the disease status from the non-disease status is required.

Typically, in evaluating a diagnostic test's discriminatory ability, the available data come from medical records of patients who undergo the test. The accuracy of the test under study is ideally evaluated by comparison with a perfect gold standard test, which assesses disease status with certainty. In practice, however, a gold standard may be too expensive, or too invasive or both for regular use. Hence, only a subset of patients undergoes disease verification, and the decision to send a patient to verification is often based on the test result and other patient characteristics. As noted by many authors (see Begg and Greenes [1], Begg [2] and Zhou [3], among others), summary measures of test performance based on data from patients with verified disease status only may be badly biased. This bias is usually referred to as verification bias.

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. Given the cutpoint c, the sensitivity $Se(c)$ is the probability of a true positive, i.e., the probability that the test correctly identifies a

***Corresponding author: Gianfranco Adimari,** Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova 35121, Italy, E-mail: gianfranco.adimari@unipd.it

Monica Chiogna, Department of Statistical Sciences, University of Padova, Via C. Battisti 241, Padova 35121, Italy, E-mail: monica@stat.unipd.it

diseased subject. The specificity $Sp(c)$ is the probability of a true negative, i.e. the probability that the test correctly identifies a non-diseased subject. When one varies the cutpoint throughout the entire real line, the resulting pairs (1-specificity, sensitivity) form the ROC curve. A commonly used summary measure that aggregates performance information of the test is the area under the ROC curve (AUC). See, for example, Zhou et al. [4] as a general reference.

In the presence of verification bias, under the assumption that the true disease status, if missing, is missing at random (MAR), estimation of the ROC curve of a continuous test, i.e., estimation of sensitivity and specificity, has been discussed in Alonzo and Pepe [5], where alternative estimators are reviewed and compared. MAR assumption states that the probability of a subject having the disease status verified is purely determined by the test result and the subjects' observed characteristics and is conditionally independent of the unknown true disease status. This corresponds to a so-called ignorable missingness, which is often assumed in practice. Estimation of the ROC curve when the true diseased status is subject to non-ignorable missingness is tackled in Fluss et al. [6] and Liu and Zhou [7]. In all these cases, however, inference on the ROC curve requires specification of a parametric regression model for the probability of a subject being diseased and/or verified. A wrong specification of these parametric models affects the behavior of the estimators.

To reduce the effects of model misspecification, He and McDermott [8] propose a method that stratifies the verified sample into several subsamples that have homogeneous propensity scores (the conditional probabilities of verification) and allows correction for verification bias within each subsample. Parametric models are still used to estimate the propensity scores, but since the estimated propensity scores are only used for the purpose of stratification, the estimators of sensitivity and specificity are less sensitive to model misspecification. The method applies to binary tests under the MAR assumption.

In this paper, we propose a fully nonparametric method for the estimation of the ROC curve of a continuous test under verification bias. The proposed method is based on nearest-neighbor imputation and adopts generic smooth regression models for both the probability that a subject is diseased and the probability that it is verified. Our choice is motivated by the results in Ning and Cheng [9], according to which the nearest-neighbor imputation method favorably compares with other nonparametric imputation methods in estimating a population mean.

The estimators for the sensitivity and the specificity obtained by the new approach are shown to be consistent and asymptotically normal under the MAR assumption. Estimation of their variance is also discussed. Some simulation results and an illustrative example show usefulness of our proposal and advantages in comparison with known estimators.

The paper is organized as follows. In Section 2, we give a brief review of existing methods for estimating the ROC curve under verification bias. Section 3 describes the proposed approach, giving theoretical justification. Section 4 presents some results of a simulation study carried out to compare the new method with the existing methods. In Section 5, we illustrate the method with an example, and Section 6 contains details about variance estimation. A concluding discussion is given in Section 7.

2 Background

In this section, we review current bias-correction methods in the presence of verification bias, as presented in Alonzo and Pepe [5]. Let T_i denote the continuous test result from a diagnostic test, and let D_i denote the binary disease status, $i = 1, \ldots, n$, where $D_i = 1$ indicates the ith patient is diseased and $D_i = 0$ indicates the ith patient is free of disease. Let V_i denote the binary verification status of the ith patients, with $V_i = 1$ if the ith patient has the true disease status verified, and $V_i = 0$ otherwise. In practice, some information, other than the results from the test, can be obtained for each patient. Let X_i be a vector of observed covariates for the ith patient that may be associated with both D_i and V_i.

When all patients are verified, i.e., $V_i = 1, i = 1, \ldots, n$, a complete data set is obtained. In this case, for any cutpoint c, the sensitivity $Se(c)$ and the specificity $Sp(c)$ could be easily estimated by

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

where $I(\cdot)$ is the indicator function. $\widehat{Se}(c)$ and $\widehat{Sp}(c)$ are unbiased estimators for $Se(c)$ and $Sp(c)$, respectively.

If not all patients have their disease status verified, several estimators based on the MAR assumption have been proposed. MAR assumption states that the binary responses D and V are mutually independent given the test result T and the covariates X, i.e.,

$$\Pr(V = 1|D, T, X) = \Pr(V = 1|T, X). \tag{1}$$

The so-called full imputation (FI) estimators of $Se(c)$ and $Sp(c)$ are

$$\widehat{Se}_{FI}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\hat{\rho}_i}{\sum_{i=1}^{n} \hat{\rho}_i}, \quad \widehat{Sp}_{FI}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)(1 - \hat{\rho}_i)}{\sum_{i=1}^{n} (1 - \hat{\rho}_i)}.$$

Parametric models, such as logistic regression models, have to be used to obtain the estimate $\hat{\rho}_i$ of $\rho_i = \Pr(D_i = 1|T_i, X_i)$ using only data from verified subjects. Mean score imputation (MSI) is another possible approach that only imputes disease status for subjects who are not in the verification sample. In this case,

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

$$\widehat{Sp}_{MSI}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)\{V_i(1 - D_i) + (1 - V_i)(1 - \hat{\rho}_i)\}}{\sum_{i=1}^{n} \{V_i(1 - D_i) + (1 - V_i)(1 - \hat{\rho}_i)\}}.$$

The inverse probability weighting (IPW) estimator weights each verified subject by the inverse of the probability that the subject is selected for verification. Therefore, the estimators of $Se(c)$ and $Sp(c)$ are

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

$$\widehat{Sp}_{IPW}(c) = \frac{\sum_{i=1}^{n} I(T_i < c) V_i (1 - D_i) \hat{\pi}_i^{-1}}{\sum_{i=1}^{n} V_i (1 - D_i) \hat{\pi}_i^{-1}},$$

where $\hat{\pi}_i$ is an estimate of $\pi_i = \Pr(V_i = 1|T_i, X_i)$. Finally, the semiparametric efficient (SPE) estimators are

$$\widehat{Se}_{SPE}(c) = \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (\hat{\pi}_i - V_i)\hat{\rho}_i\}\hat{\pi}_i^{-1}}{\sum_{i=1}^{n} \{V_i D_i + (\hat{\pi}_i - V_i)\hat{\rho}_i\}\hat{\pi}_i^{-1}},$$

$$\widehat{Se}_{SPE}(c) = \frac{\sum_{i=1}^{n} I(T_i < c)\{V_i(1 - D_i) + (\hat{\pi}_i - V_i)(1 - \hat{\rho}_i)\}\hat{\pi}_i^{-1}}{\sum_{i=1}^{n} \{V_i(1 - D_i) + (\hat{\pi}_i - V_i)(1 - \hat{\rho}_i)\}\hat{\pi}_i^{-1}}.$$

Alonzo and Pepe [5] find that SPE estimators are doubly robust in the sense that they are consistent if either π_i's or ρ_i's are estimated consistently.

3 The proposal

All the verification bias-corrected estimators of $Se(c)$ and $Sp(c)$ reviewed in the previous section require a regression model to be fitted for a binary response, D or V. The FI and MSI approaches require estimates of ρ_i's, whereas the IPW approach requires estimates of π_i's. The SPE approach requires estimates of both ρ_i's and π_i,'s although only one of the two sets of probabilities needs to be estimated consistently. Typically, suitable generalized linear regression models are employed to this end. However, a wrong specification of such parametric models might strongly affect the behavior of the estimators.

To avoid misspecification problems, in what follows we propose a fully nonparametric approach to the estimation of $Se(c)$ and $Sp(c)$. Our approach is based on the K-nearest-neighbor (KNN) imputation estimator of the mean of a response variable as discussed in Ning and Cheng [9].

Hereafter, we will assume $Y = (T, X)^{\mathrm{T}}$ to be a continuous-valued random vector. Let θ_1 be the disease prevalence, i.e., $\theta_1 = E(D) = \Pr(D = 1)$. As θ_1 is a mean, following Ning and Cheng [9], for a finite positive integer K and a suitable distance measure, a nearest-neighbor imputation estimator of θ_1, based on the sample $(Y_i, D_i, V_i), i = 1, \ldots, n,$ may be defined as

$$\hat{\theta}_1 = \frac{1}{n}\sum_{i=1}^{n}\{V_i D_i + (1 - V_i)\hat{\rho}_{Ki}\}, \tag{2}$$

where $\hat{\rho}_{Ki} = \frac{1}{K}\sum_{j=1}^{K} D_{i(j)}$, and $\{(Y_{i(j)}, D_{i(j)}) : V_{i(j)} = 1, j = 1, \ldots, K\}$ is a set of K observed data pairs and $Y_{i(j)}$ denotes the jth nearest neighbor to $Y_i = (T_i, X_i)^{\mathrm{T}}$ among all Y's corresponding to the verified patients, i.e., to those D_h's with $V_h = 1$.

Let $\theta_2 = \Pr(T \geq c, D = 1)$ and $\theta_3 = \Pr(T \geq c, D = 0)$. Then $Se(c) = \frac{\theta_2}{\theta_1}$ and $Sp(c) = 1 - \frac{\theta_3}{1-\theta_1}$. Similarly to $\hat{\theta}_1$, KNN estimators for θ_2 and θ_3 can be defined as:

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (1 - V_i)\hat{\rho}_{Ki}\},$$

$$\hat{\theta}_3 = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)\{V_i(1 - D_i) + (1 - V_i)(1 - \hat{\rho}_{Ki})\}.$$

Therefore

$$\widehat{Se}_{KNN}(c) = \frac{\hat{\theta}_2}{\hat{\theta}_1} \quad \textit{and} \quad \widehat{Sp}_{KNN}(c) = \frac{1 - \hat{\theta}_1 - \hat{\theta}_3}{1 - \hat{\theta}_1},$$

are KNN imputation estimators for the sensitivity $Se(c)$ and the specificity $Sp(c)$, respectively. The following theorem gives asymptotic normality of $\widehat{Se}_{KNN}(c)$ and $\widehat{Sp}_{KNN}(c)$

Let $\rho(y) = \Pr(D = 1|Y = y)$ and $\pi(y) = \Pr(V = 1|Y = y)$.

Theorem 1 *Assume (1) and first-order differentiability of the functions $\rho(y)$ and $\pi(y)$. Moreover, assume that $E(1/\pi(Y)) < \infty$. Then, for a fixed cutpoint c, the KNN imputation estimators $\widehat{Se}_{KNN}(c)$ and $\widehat{Sp}_{KNN}(c)$ based on the sample $(Y_i, D_i, V_i), i = 1, \ldots, n,$ are consistent and asymptotically normally distributed.*

Proof 1 *Since $E(D^2) < \infty$, $Var(D|Y = y) = \rho(y)(1 - \rho(y)) < \infty$, $\rho(y)$ and $\pi(y)$ are finite and first order differentiable, by Theorem 1 in Ning and Cheng [9], the KNN imputation estimator $\hat{\theta}_1$ is consistent and asymptotically normally distributed, that is*

$$\sqrt{n}(\hat{\theta}_1 - \theta_1) \to N(0, \sigma_1^2), \tag{3}$$

as n goes to infinity, where

$$\sigma_1^2 = \theta_1(1-\theta_1) + E[\rho(Y)(1-\rho(Y))(1-\pi(Y))]\left(1+\frac{1}{K}\right) + E\left[\frac{\rho(Y)(1-\rho(Y))(1-\pi(Y))^2}{\pi(Y)}\right].$$

Moreover, one can write

$$\hat{\theta}_2 = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (1-V_i)\rho_i\} + \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)(1-V_i)(\hat{\rho}_{Ki} - \rho_i), \tag{4}$$

and

$$\hat{\theta}_3 = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)\{V_i(1-D_i) + (1-V_i)(1-\rho_i)\} - \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)(1-V_i)(\hat{\rho}_{Ki} - \rho_i). \tag{5}$$

Hence, conditions stated in the theorem allow to apply the arguments given in the proof of Theorem 1 in Ning and Cheng [9] showing, in particular, that

$$\frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)(1-V_i)(\hat{\rho}_{Ki} - \rho_i) = W + o_p(n^{-1/2}),$$

where $W = \frac{1}{n}\sum_{i=1}^{n} I(T_i \geq c)(1-V_i)\left[\frac{1}{K}\sum_{j=1}^{K}\left(V_{i(j)}D_{i(j)} - \rho_{i(j)}\right)\right]$, and

$$\sqrt{n}W \to N\left(0, \frac{1}{K}E\left[(1-\pi(Y))\sigma^2(Y)\right] + E\left[\frac{(1-\pi(Y))^2\sigma^2(Y)}{\pi(Y)}\right]\right),$$

in distribution (here, $\sigma^2(Y)$ *denotes the conditional variance of* $I(T \geq c, D = 1)$ *given* Y*). Furthermore,* $\sqrt{n}W$ *behaves asymptotically as a sample mean. This, together with an application of the standard central limit theorem to the first term of the right hand side of equations (4) and (5), leads to asymptotic results for* $\hat{\theta}_2$ *and* $\hat{\theta}_3$ *similar to that in (3). That is*

$$\sqrt{n}(\hat{\theta}_2 - \theta_2) \to N(0, \sigma_2^2), \qquad \sqrt{n}(\hat{\theta}_3 - \theta_3) \to N(0, \sigma_3^2),$$

as n *goes to infinity, for suitable values* σ_2^2 *and* σ_3^2 *(see also Section 6). Finally,* $\hat{\theta}_1$, $\hat{\theta}_2$ *and* $\hat{\theta}_3$ *are jointly asymptotically normal. Thus, by a standard application of the delta method,* $\widehat{Se}_{KNN}(c) = \frac{\hat{\theta}_2}{\hat{\theta}_1}$ *and* $\widehat{Sp}_{KNN}(c) = \frac{1-\hat{\theta}_1-\hat{\theta}_3}{1-\hat{\theta}_1}$ *are consistent and asymptotically normal estimators of* $Se(c)$ *and* $Sp(c)$*, respectively.*

It is straightforward to show that estimators $\widehat{Se}_{KNN}(c)$ and $\widehat{Sp}_{KNN}(c)$ are nonparametric version of the MSI estimators, i.e.,

$$\begin{aligned}
\widehat{Se}_{KNN}(c) &= \frac{\sum_{i=1}^{n} I(T_i \geq c)\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}}{\sum_{i=1}^{n}\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}}, \\
\widehat{Sp}_{KNN}(c) &= \frac{\sum_{i=1}^{n} I(T_i < c)\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{Ki})\}}{\sum_{i=1}^{n}\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{Ki})\}}.
\end{aligned} \tag{6}$$

Clearly, by varying c, the pairs $(1 - \widehat{Sp}_{KNN}(c), \widehat{Se}_{KNN}(c))$ give rise to the nonparametric verification bias-corrected estimate of the ROC curve. Moreover, it is worth noting that (2) gives a fully nonparametric estimator for the disease prevalence that is alternative to the estimators, obtained by the FI, MSI, IPW and SPE methods, discussed in Alonzo and Pepe [10].

In practice, the use of our estimators requires to select the neighborhood size K and a suitable distance measure. Such aspects, touched upon in the following section, are discussed in Section S1 of Supplementary Material.

4 Simulation study

In this section, Monte Carlo experiments are used to compare the new method with existing approaches with respect to bias and standard deviation. In particular, we compare the ability of the MSI, IPW, SPE and KNN methods to estimate the sensitivity and the specificity of a test. We do not consider the FI method because of its similarities with the MSI method. As for the KNN method, we give the results for the estimators based on the quite commonly used Euclidean distance and on values of K equal to 1 and 3. This choice is supported by the results of a preliminary simulation study, in which KNN estimators based on various distance measures (Manhattan, Euclidean, Lagrange and Mahalanobis) and on different neighborhood sizes ($K = 1,\ 3,\ 5,\ 10,\ 20$) have been compared (see Supplementary Material, Section S1).

From Section 2, the MSI method requires a parametric model for $\rho(y)$, the IPW method requires a parametric model for $\pi(y)$, and the SPE method requires both models. A wrong specification of such models may affect the est imation. Hence, in the simulation study we consider two scenarios: (i) the models for $\rho(y)$ and $\pi(y)$ are both correctly specified, (ii) the models for $\rho(y)$ and $\pi(y)$ are both misspecified. Scenario (i) allows to evaluate the behavior of the proposed estimators in samples of moderate sizes, where the MSI, IPW and SPE estimators are expected to well behave. On the other side, scenario (ii) allows to look for weaknesses of existing methods and to highlight the potential advantages of the new proposal.

Simulation settings are similar to those in Alonzo and Pepe [5] and He and McDermott [8]. Starting from two independent random variables $Z_1 \sim N(0, 0.5)$ and $Z_2 \sim N(0, 0.5)$, the disease indicator D is specified as $D = I[g(Z_1, Z_2) > r_1]$. The threshold r_1 determines the disease prevalence (in what follows, we choose r_1 to make the disease prevalence 0.25) and different specifications of the function $g(Z_1, Z_2)$ give rise to different disease processes. The diagnostic test result T and an auxiliary covariate X are generated to be related to D through Z_1 and Z_2. More precisely, $T = h(Z_1, Z_2) + \varepsilon_1$ and $X = f(Z_1, Z_2) + \varepsilon_2$, for suitable functions $h(\cdot, \cdot)$ and $f(\cdot, \cdot)$, where ε_1 and ε_2 are independent $N(0, 0.25)$ random variables, independent also from Z_1 and Z_2. Finally, the verification probability π is set to be a suitable function of T and X, in accordance with the MAR assumption. The number of replicates in each simulation experiment is 5,000.

(i) Models for $\rho(y)$ and $\pi(y)$ both correctly specified.

We set $g(Z_1, Z_2) = f(Z_1, Z_2) = Z_1 + Z_2$, $h(Z_1, Z_2) = \alpha(Z_1 + Z_2)$, and $\pi(T, X) = \dfrac{e^{\delta_0 + \delta_1 T + \delta_2 X}}{1 + e^{\delta_0 + \delta_1 T + \delta_2 X}}$. We fix $\delta_0 = 0.05$, $\delta_1 = 0.9$, $\delta_2 = 0.7$. This choice corresponds to a verification rate of about 0.51. As for α, we choose three different values, i.e., 0.5, 1 and 1.5 that give rise to different variances of T, as well as to different correlations between T and X, with larger values giving rise to higher variances and correlations. In particular, on going from $\alpha = 0.5$ to $\alpha = 1.5$ the variance of T becomes five times greater. Moreover, we consider four values for the cutpoint c, i.e., 0.2, 0.5, 0.8, and 1.2. Obviously, each combination (α, c) determines a different true value for the pair (sensitivity, specificity), given by

$$\left(Se(c) = 1 - \frac{\int_{r_1}^{+\infty} \Phi(\frac{c - \alpha z}{\sqrt{0.25}})\varphi(z)dz}{1 - \Phi(r_1)}, \qquad Sp(c) = \frac{\int_{-\infty}^{r_1} \Phi(\frac{c - \alpha z}{\sqrt{0.25}})\varphi(z)dz}{\Phi(r_1)} \right),$$

where $\varphi(\cdot)$ and $\Phi(\cdot)$ are the density function and the cumulative distribution function of the standard normal random variable, respectively.

According to the aim of the study in this scenario, we fix two sample sizes, a relatively small one, i.e., $n = 50$, and a moderate one, i.e., $n = 100$. This allows to evaluate the behavior of the proposed estimators in settings where the MSI, IPW and SPE estimators are expected to well behave.

To estimate the conditional disease probabilities, we use a generalized linear model for D given T and X with probit link; this model is correctly specified (see Alonzo and Pepe [5]). The conditional verification probabilities are estimated from a logistic regression model with V as the response and T and X as predictors. Evidently, also this model is correct.

Tables 1 and 2 show Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity. Results concern the estimators IPW, MSI, SPE and the new proposals 1NN and 3NN, i.e., the KNN estimator with $K = 1$ and $K = 3$, respectively, computed using the Euclidean distance. From the simulation results it is clear that all of the methods behave well if both parametric models for $\rho(y)$ and $\pi(y)$ are correctly specified, with the IPW method showing slightly poorer performances in some circumstances. In terms of bias and standard deviation, the new proposals compare very well with existing estimators. Moreover, the estimators 1NN and 3NN seem to achieve similar performances, making the choice of the number K of nearest neighbors not particularly crucial (within the range 1–3).

Tables 1 and 2 allow also to gain insight into the effect on results of different variances of T and of different correlations between T and X. By crossing values of α and c giving rise to comparable values of sensitivity or specificity, it is possible to note that, for all considered estimators, the obtained Monte Carlo means and standard deviations are essentially not influenced by the different values of variance and correlation. As far as sensitivity is concerned, for example, one can compare results obtained for $\alpha = 0.5$ and $c = 0.2$ (true sensitivity equal to 0.782), with results obtained for $\alpha = 1.5$ and $c = 1.2$ (true sensitivity equal to 0.784). As for specificity, one can compare results obtained for $\alpha = 0.5$ and $c = 0.2$ (true specificity equal to 0.742), with results obtained for $\alpha = 1$ and $c = 0.2$ (true specificity equal to 0.745) or $\alpha = 1.5$ and $c = 0.2$ (true specificity equal to 0.731).

As pointed out by a Referee, the values chosen for α in Tables 1 and 2 refer to situations where the diagnostic tests perform well, i.e., situations where the true AUC value ranges from 0.85 to 0.97. Performance of estimators in situations where the true AUC value of the test is relatively small (0.59 and 0.71) are given in Section S2, Supplementary Material. Results show the same behavior as the one shown in Tables 1 and 2.

Simulation results allowing to explore the effect of a multidimensional vector of auxiliary covariates are given in Section S4, Supplementary Material. A vector X of dimension 3 is employed. Compared with results in Tables 1 and 2, results in Tables 10 and 11, Supplementary Material, show some loss of efficiency of the KNN estimators with respect to the parametric competitors.

(ii) Models for $\rho(y)$ and $\pi(y)$ both misspecified.

We set $g(Z_1, Z_2) = \exp[2(Z_1 Z_2)^2]$, $h(Z_1, Z_2) = 2(Z_1 Z_2)^2$, $f(Z_1, Z_2) = \sqrt{2}(Z_1^2 + Z_2^2)$, and $\pi(T, X) = 0.05 + \delta I[T > 1.2] + (1 - 0.05 - \delta) I[X > 1.95]$. In this case, the verification probabilities are: 1 for those subjects with $T > 1.2$ and $X > 1.95$; $1 - \delta$ for those subjects with $T \leq 1.2$ and $X > 1.95$; $0.05 + \delta$ for those subjects with $T > 1.2$ and $X \leq 1.95$; 0.05 otherwise. The values 1.2 and 1.95 correspond roughly to the 92-th and the 86-th percentile of the distributions of T and X, respectively. The value of δ is allowed to range from 0.1 to 0.9 with steps of 0.2. By varying δ, one can vary the strength of the dependence among V, T, and X: small values of δ indicate a strong dependence of V on X, whereas high values of δ indicate a strong dependence of V on T. Finally, for the cutpoint c, we choose three different values, i.e., 0.2, 0.4, 0.6, that give rise to three different values for the target pair (sensitivity, specificity). The aim in this scenario is to compare the estimators when the complete data set provides a great amount of information, in order to highlight possible weaknesses of competitors of our KNN estimators, in particular their possible inconsistency. Therefore, the required size for generating samples should be high enough to guarantee both reliable estimates from the complete data set and a sufficiently high number of verified healthy and diseased

Table 1: Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity, when the models for $\rho(y)$ and $\pi(y)$ are correctly specified. "True" denotes the true parameter value. Sample size = 50.

	$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
$\alpha = 0.5$								
Sensitivity								
True	0.782		0.590		0.377		0.154	
IPW	0.787	(0.146)	0.598	(0.169)	0.384	(0.157)	0.163	(0.116)
MSI	0.783	(0.133)	0.595	(0.158)	0.383	(0.147)	0.161	(0.109)
SPE	0.783	(0.140)	0.595	(0.161)	0.382	(0.149)	0.162	(0.110)
1NN	0.783	(0.143)	0.596	(0.166)	0.382	(0.153)	0.162	(0.112)
3NN	0.775	(0.136)	0.587	(0.159)	0.376	(0.147)	0.159	(0.109)
Specificity								
True	0.742		0.877		0.953		0.992	
IPW	0.735	(0.100)	0.873	(0.070)	0.953	(0.043)	0.992	(0.017)
MSI	0.742	(0.074)	0.877	(0.057)	0.955	(0.035)	0.993	(0.014)
SPE	0.742	(0.075)	0.876	(0.058)	0.954	(0.036)	0.993	(0.015)
1NN	0.742	(0.076)	0.877	(0.058)	0.954	(0.036)	0.992	(0.015)
3NN	0.741	(0.074)	0.875	(0.057)	0.954	(0.035)	0.992	(0.015)
$\alpha = 1$								
Sensitivity								
True	0.951		0.874		0.742		0.513	
IPW	0.954	(0.077)	0.877	(0.117)	0.746	(0.148)	0.520	(0.159)
MSI	0.951	(0.067)	0.875	(0.110)	0.747	(0.140)	0.516	(0.152)
SPE	0.953	(0.076)	0.876	(0.118)	0.746	(0.142)	0.516	(0.153)
1NN	0.950	(0.079)	0.872	(0.117)	0.744	(0.142)	0.515	(0.156)
3NN	0.944	(0.077)	0.863	(0.117)	0.735	(0.142)	0.509	(0.153)
Specificity								
True	0.745		0.855		0.931		0.982	
IPW	0.729	(0.105)	0.846	(0.078)	0.926	(0.052)	0.980	(0.027)
MSI	0.745	(0.074)	0.856	(0.060)	0.931	(0.043)	0.982	(0.022)
SPE	0.745	(0.074)	0.856	(0.061)	0.931	(0.044)	0.982	(0.023)
1NN	0.745	(0.075)	0.856	(0.062)	0.931	(0.044)	0.981	(0.024)
3NN	0.744	(0.074)	0.855	(0.060)	0.930	(0.043)	0.981	(0.023)
$\alpha = 1.5$								
Sensitivity								
True	0.991		0.969		0.918		0.784	
IPW	0.992	(0.031)	0.973	(0.058)	0.920	(0.092)	0.787	(0.133)
MSI	0.990	(0.028)	0.970	(0.053)	0.920	(0.086)	0.786	(0.130)
SPE	0.991	(0.032)	0.971	(0.057)	0.920	(0.089)	0.785	(0.131)
1NN	0.990	(0.033)	0.970	(0.060)	0.918	(0.092)	0.783	(0.134)
3NN	0.986	(0.038)	0.965	(0.061)	0.912	(0.091)	0.776	(0.133)
Specificity								
True	0.731		0.822		0.897		0.963	
IPW	0.700	(0.121)	0.803	(0.092)	0.886	(0.069)	0.958	(0.040)
MSI	0.730	(0.076)	0.824	(0.065)	0.898	(0.053)	0.963	(0.032)
SPE	0.731	(0.076)	0.824	(0.066)	0.898	(0.053)	0.963	(0.033)
1NN	0.731	(0.077)	0.824	(0.067)	0.898	(0.054)	0.962	(0.033)
3NN	0.731	(0.076)	0.824	(0.065)	0.897	(0.053)	0.962	(0.032)

Table 2: Monte Carlo means and standard deviations (in brackets) of the estimators for the sensitivity and the specificity, when the models for $\rho(y)$ and $\pi(y)$ are correctly specified. "True" denotes the true parameter value. Sample size = 100.

	$c = 0.2$		$c = 0.5$		$c = 0.8$		$c = 1.2$	
$\alpha = 0.5$								
Sensitivity								
True	0.782		0.590		0.377		0.154	
IPW	0.785	(0.102)	0.595	(0.116)	0.379	(0.109)	0.159	(0.079)
MSI	0.785	(0.091)	0.595	(0.107)	0.380	(0.103)	0.159	(0.076)
SPE	0.785	(0.097)	0.594	(0.110)	0.379	(0.104)	0.159	(0.076)
1NN	0.783	(0.101)	0.594	(0.115)	0.378	(0.106)	0.159	(0.077)
3NN	0.780	(0.096)	0.590	(0.110)	0.376	(0.104)	0.158	(0.075)
Specificity								
True	0.742		0.877		0.953		0.992	
IPW	0.738	(0.068)	0.877	(0.047)	0.954	(0.029)	0.992	(0.012)
MSI	0.742	(0.052)	0.878	(0.038)	0.955	(0.025)	0.992	(0.010)
SPE	0.742	(0.053)	0.878	(0.039)	0.954	(0.025)	0.992	(0.011)
1NN	0.742	(0.054)	0.878	(0.040)	0.954	(0.026)	0.992	(0.011)
3NN	0.741	(0.053)	0.877	(0.039)	0.954	(0.025)	0.992	(0.011)
$\alpha = 1$								
Sensitivity								
True	0.951		0.874		0.742		0.513	
IPW	0.952	(0.054)	0.875	(0.081)	0.746	(0.102)	0.517	(0.112)
MSI	0.950	(0.046)	0.875	(0.074)	0.746	(0.096)	0.516	(0.108)
SPE	0.951	(0.053)	0.875	(0.078)	0.746	(0.099)	0.516	(0.109)
1NN	0.950	(0.056)	0.873	(0.083)	0.744	(0.103)	0.515	(0.111)
3NN	0.947	(0.053)	0.870	(0.079)	0.741	(0.099)	0.512	(0.109)
Specificity								
True	0.745		0.855		0.931		0.982	
IPW	0.738	(0.073)	0.851	(0.052)	0.929	(0.036)	0.982	(0.018)
MSI	0.745	(0.052)	0.855	(0.042)	0.931	(0.030)	0.982	(0.015)
SPE	0.746	(0.053)	0.855	(0.043)	0.931	(0.031)	0.982	(0.016)
1NN	0.745	(0.054)	0.855	(0.044)	0.931	(0.032)	0.982	(0.016)
3NN	0.745	(0.053)	0.855	(0.043)	0.931	(0.031)	0.982	(0.016)
$\alpha = 1.5$								
Sensitivity								
True	0.991		0.969		0.918		0.784	
IPW	0.992	(0.022)	0.970	(0.041)	0.918	(0.064)	0.785	(0.093)
MSI	0.991	(0.018)	0.969	(0.036)	0.918	(0.059)	0.785	(0.090)
SPE	0.992	(0.022)	0.970	(0.040)	0.918	(0.063)	0.784	(0.091)
1NN	0.991	(0.024)	0.969	(0.043)	0.917	(0.066)	0.783	(0.094)
3NN	0.990	(0.022)	0.967	(0.041)	0.914	(0.063)	0.780	(0.091)
Specificity								
True	0.731		0.822		0.897		0.963	
IPW	0.713	(0.082)	0.812	(0.062)	0.893	(0.046)	0.961	(0.026)
MSI	0.731	(0.052)	0.823	(0.045)	0.898	(0.036)	0.963	(0.022)
SPE	0.731	(0.052)	0.823	(0.046)	0.898	(0.037)	0.963	(0.023)
1NN	0.731	(0.053)	0.824	(0.047)	0.898	(0.037)	0.963	(0.023)
3NN	0.731	(0.053)	0.823	(0.046)	0.898	(0.037)	0.963	(0.022)

subjects. In the setting of scenario (ii), for δ going from 0.1 to 0.9, the verification rate ranges roughly from 0.29 to 0.18 and, within healthy subjects, from 0.11 to 0.05. This has led us to the choice of $n = 1000$.

To estimate the conditional disease probabilities, we use a generalized linear model for D given T and X with logit link; this model is misspecified. The conditional verification probabilities are estimated from a logistic regression model with V as the response and T as predictor. Clearly, also this model is misspecified.

Table 3 presents Monte Carlo means and standard deviations (across 5,000 replications) for the estimators of the sensitivity and the specificity. Results concern the estimators IPW, MSI, SPE, 1NN and 3NN. Moreover, results for the estimators based on complete data (denoted by "Full" in the table), that is with all cases verified, are also presented. Given the large sample size utilized in this setting, we expect that the Monte Carlo means for the Full estimators represent a good approximation of the true values of the sensitivity and the specificity and they are therefore used as the benchmark values.

Table 3 clearly shows limitations of parametric estimators when models for $\rho(y)$ and $\pi(y)$ are misspecified. In particular, in terms of bias, the IPW, MSI and SPE methods perform almost always poorly, with high distortion in some cases. Moreover, the Monte Carlo standard deviations shown in the table indicate that the SPE method (and sometime also the IPW method) might yield very unstable estimates. In fact, the SPE estimates may even fall outside the interval (0,1). In our simulations, in the worst case this event happened about 20 times across 5,000 replications.

Overall, the new estimators 1NN and 3NN perform well in terms of both bias and standard deviation. In particular, they yield estimates that are, in all cases, close to the full data estimates (see also results in Section S3, Supplementary Material, where some simulations have been produced for a smaller sample size). The estimator 3NN appears to be slightly more biased than 1NN, but, on the other side, with slightly less variance. Note that in this setting the function $\pi(y)$ used for the verification process is not smooth. Then, the KNN estimators seem to show also some degree of robustness against violation of smoothness assumptions. This is not surprising because, as stated in Section 2 of Ning and Cheng [9], "the NN rule is basically unaffected by discontinuity of $\pi(y)$, sparse data or multi-dimensional covariate".

5 An illustration

To illustrate the application of the method developed in the previous sections, we utilize the Wisconsin Breast Cancer Data, publicly available at the UCI Machine Learning Repository [11]. The construction of the dataset was motivated by the need to accurately diagnose breast masses on the basis, solely, of a Fine Needle Aspiration (FNA). The dataset collects various features which are computed from a digitized image of a FNA of a breast mass, describing characteristics of the cell nuclei present in the image. A total of 30 nuclear features are computed on each of 569 samples, of which 357 are benign and 212 malignant. The dataset has been extensively used in the literature. The interested reader can refer to the UCI Machine Learning Repository documentation for retrieving information about the dataset creation, the description of its attributes, and a list of relevant papers using or citing this data set.

Here, we use one of the features, i.e., the worst radius (WR), as the test to diagnose malignant breast masses, and one of the remaining features, i.e., the worst concave point (WCP), as a covariate giving auxiliary information. Our aim is to estimate the ROC curve of the test WR. To mimic verification bias, a subset of the complete dataset is constructed. In this subset, the test WR and the covariate WCP are known for all samples, but the true status (benign or malignant) is available only for some samples, that we select according to the following mechanism. We select all samples having a value for both WR and WCP above their respective medians; we do not select samples having a value for both WR and WCP below their respective medians; we select all remaining samples with probability equal to 0.95.

The obtained dataset shows a percentage of samples with true status known (verified) of about 58%. The percentage of benign samples is about 36% among verified samples and 99% among non-verified samples.

Table 3: Mean estimated sensitivity, mean estimated specificity and standard deviation (in brackets) from 5,000 replications when both models for $\rho(y)$ and $\pi(y)$ are misspecified and the cutpoint c is set equal to 0.2, 0.4, 0.6. "Full" indicates the estimator based on complete data, which does not change with δ. Sample size = 1,000.

δ	IPW	MSI	SPE	1NN	3NN	Full
Sensitivity						
$c = 0.2$						
0.1	0.778 (0.052)	0.868 (0.029)	0.846 (0.051)	0.888 (0.055)	0.885 (0.047)	
0.3	0.767 (0.060)	0.877 (0.029)	0.858 (0.061)	0.889 (0.054)	0.885 (0.048)	
0.5	0.752 (0.077)	0.887 (0.030)	0.870 (0.083)	0.886 (0.057)	0.882 (0.049)	0.888 (0.020)
0.7	0.736 (0.108)	0.898 (0.032)	0.893 (0.158)	0.886 (0.060)	0.880 (0.052)	
0.9	0.744 (0.169)	0.903 (0.040)	0.929 (0.731)	0.879 (0.067)	0.871 (0.058)	
$c = 0.4$						
0.1	0.685 (0.051)	0.778 (0.039)	0.759 (0.050)	0.818 (0.064)	0.814 (0.056)	
0.3	0.679 (0.060)	0.794 (0.040)	0.778 (0.060)	0.821 (0.065)	0.816 (0.056)	
0.5	0.671 (0.074)	0.810 (0.041)	0.796 (0.077)	0.821 (0.066)	0.816 (0.058)	0.820 (0.024)
0.7	0.663 (0.100)	0.828 (0.043)	0.819 (0.250)	0.820 (0.068)	0.813 (0.060)	
0.9	0.684 (0.164)	0.839 (0.056)	0.908 (3.428)	0.813 (0.080)	0.803 (0.070)	
$c = 0.6$						
0.1	0.593 (0.049)	0.672 (0.046)	0.658 (0.051)	0.738 (0.069)	0.734 (0.061)	
0.3	0.590 (0.056)	0.691 (0.047)	0.678 (0.058)	0.739 (0.069)	0.734 (0.061)	
0.5	0.588 (0.070)	0.713 (0.049)	0.701 (0.072)	0.737 (0.071)	0.732 (0.063)	0.737 (0.028)
0.7	0.593 (0.093)	0.739 (0.053)	0.733 (0.103)	0.738 (0.074)	0.731 (0.066)	
0.9	0.638 (0.153)	0.759 (0.066)	0.867 (3.715)	0.732 (0.084)	0.724 (0.074)	
Specificity						
$c = 0.2$						
0.1	0.776 (0.031)	0.649 (0.024)	0.641 (0.028)	0.603 (0.026)	0.603 (0.024)	
0.3	0.799 (0.032)	0.637 (0.023)	0.630 (0.028)	0.603 (0.026)	0.603 (0.024)	
0.5	0.827 (0.032)	0.626 (0.023)	0.620 (0.031)	0.603 (0.027)	0.603 (0.024)	0.602 (0.018)
0.7	0.865 (0.033)	0.614 (0.021)	0.607 (0.096)	0.603 (0.028)	0.603 (0.025)	
0.9	0.914 (0.030)	0.604 (0.021)	0.613 (0.661)	0.602 (0.029)	0.602 (0.026)	
$c = 0.4$						
0.1	0.870 (0.022)	0.788 (0.021)	0.782 (0.024)	0.742 (0.025)	0.742 (0.022)	
0.3	0.885 (0.022)	0.776 (0.021)	0.772 (0.024)	0.743 (0.025)	0.743 (0.022)	
0.5	0.903 (0.021)	0.765 (0.020)	0.762 (0.024)	0.743 (0.025)	0.743 (0.022)	0.743 (0.016)
0.7	0.926 (0.019)	0.755 (0.020)	0.751 (0.028)	0.743 (0.025)	0.743 (0.022)	
0.9	0.956 (0.016)	0.745 (0.019)	0.739 (0.221)	0.743 (0.027)	0.744 (0.023)	
$c = 0.6$						
0.1	0.932 (0.015)	0.887 (0.016)	0.885 (0.018)	0.852 (0.022)	0.852 (0.018)	
0.3	0.939 (0.014)	0.879 (0.016)	0.876 (0.018)	0.852 (0.021)	0.852 (0.018)	
0.5	0.948 (0.013)	0.870 (0.016)	0.868 (0.018)	0.852 (0.020)	0.853 (0.018)	0.852 (0.013)
0.7	0.961 (0.011)	0.862 (0.016)	0.860 (0.019)	0.852 (0.021)	0.853 (0.018)	
0.9	0.976 (0.009)	0.854 (0.015)	0.851 (0.055)	0.852 (0.022)	0.854 (0.018)	

As traditional methods (MSI, IPW, SPE) require the use of parametric regression models for the conditional probability of a sample being malignant and/or selected (i.e., with the true status known), we use a generalized linear model for the status given WR and WCP with probit link to estimate the conditional disease probabilities, and a logistic regression model with WR and WCP as predictors to estimate the conditional selection probabilities. Clearly, this last model is misspecified.

Figure 1 shows the estimated ROC curves of the test WR obtained with the IPW, MSI, SPE, 1NN and 3NN methods. Such curves are benchmarked with the estimated ROC curve obtained from the complete dataset by using the Full estimator of sensitivity and specificity. The plot shows that the estimators MSI, 1NN and 3NN well behave, whereas the estimators IPW and SPE are highly biased. This could imply that, in our data, the probit model is a good approximation of the disease process, whereas misspecification of the selection model seems to highly affect the estimators IPW and SPE. This is somehow surprising, as far as the doubly robust SPE estimator is concerned, especially taking into account the good behavior of the MSI estimator. It is worth noting, however, that the SPE estimator produces estimates that are outside the range $(0,1)$ and estimates of the specificity around 0.82 are not monotonically increasing for increasing values of the cutpoint c, as shown in Table 4, which reports estimates obtained for the sensitivity and the specificity at some values of c.

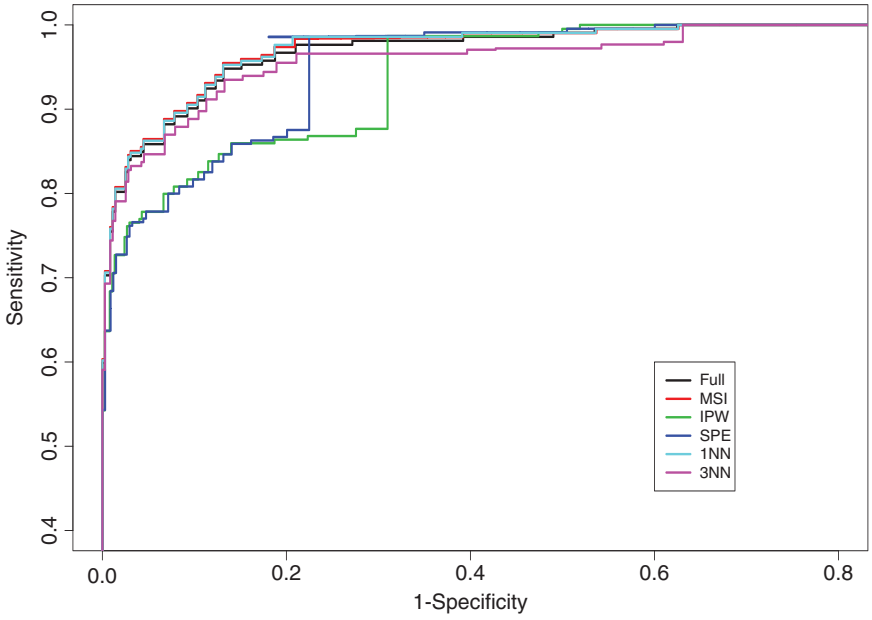

Figure 1: Illustrative example: estimated ROC curves of the test WR.

Table 4: Illustrative example: estimates for the pair (sensitivity, specificity) of the test WR obtained by the various methods at different values of the cutpoint c.

c	Full	MSI	IPW	SPE	1NN	3NN
14.569	(0.976, 0.728)	(0.984, 0.730)	(0.987, 0.558)	(0.986, 0.781)	(0.986, 0.732)	(0.966, 0.728)
14.710	(0.976, 0.739)	(0.984, 0.741)	(0.987, 0.561)	(0.986, 0.793)	(0.986, 0.743)	(0.966, 0.739)
14.851	(0.976, 0.765)	(0.984, 0.766)	(0.987, 0.569)	(0.986, **0.819**)	(0.986, 0.768)	(0.966, 0.765)
14.993	(0.967, 0.790)	(0.974, 0.791)	(0.877, 0.690)	(0.875, **0.775**)	(0.976, 0.793)	(0.955, 0.789)
15.134	(0.958, 0.812)	(0.964, 0.813)	(0.868, 0.724)	(0.867, **0.799**)	(0.962, 0.813)	(0.944, 0.811)
15.275	(0.953, 0.826)	(0.960, 0.827)	(0.864, 0.777)	(0.863, **0.814**)	(0.957, 0.827)	(0.940, 0.825)
15.417	(0.948, 0.849)	(0.955, 0.849)	(0.860, 0.813)	(0.859, 0.838)	(0.953, 0.849)	(0.935, 0.847)
15.558	(0.934, 0.868)	(0.941, 0.869)	(0.847, 0.860)	(0.846, 0.859)	(0.938, 0.869)	(0.921, 0.867)

Note: Values in bold highlight the non-monotonicity of the SPE estimator.

6 Variance estimation

In this section, we describe an approach to obtain estimates of the variances of the estimators proposed in Section 3. Such estimates could be used to build confidence intervals and perform hypothesis testing.

Recall that $\hat{\theta}_1$ in (2) is the KNN imputation estimator of the disease prevalence $\theta_1 = \Pr\{D = 1\}$, and that $\hat{\theta}_2$ and $\hat{\theta}_3$ are the KNN imputation estimators of $\theta_2 = \Pr\{T \geq c, D = 1\}$ and $\theta_3 = \Pr\{T \geq c, D = 0\}$, respectively. Moreover, recall that $Y = (T, X)^\top$.

From Section 3, $\hat{\theta}_1$ has asymptotic variance $\sigma_1^2 = \theta_1(1-\theta_1) + \omega_1^2$, where

$$\omega_1^2 = E[\rho(Y)(1-\rho(Y))(1-\pi(Y))]\left(1+\frac{1}{K}\right) + E\left[\frac{\rho(Y)(1-\rho(Y))(1-\pi(Y))^2}{\pi(Y)}\right].$$

This result follows by an application of Theorem 1 in Ning and Cheng [9]. Note that, in the expression of σ_1^2, $\theta_1(1-\theta_1)$ is the variance of D, and that the term $\rho(Y)(1-\rho(Y)$ is the conditional variance of D given Y. Therefore, taking into account that $I(T \geq c)\rho(Y)(1-\rho(Y))$ is the conditional variance of $I(T \geq c, D = 1)$ given Y, the asymptotic variance of $\hat{\theta}_2$ is given by $\sigma_2^2 = \theta_2(1-\theta_2) + \omega_2^2$, where

$$\omega_2^2 = E[I(T \geq c)\rho(Y)(1-\rho(Y))(1-\pi(Y))]\left(1+\frac{1}{K}\right) + E\left[\frac{I(T \geq c)\rho(Y)(1-\rho(Y))(1-\pi(Y))^2}{\pi(Y)}\right].$$

Similarly, for the asymptotic variance of $\hat{\theta}_3$ one obtains $\sigma_3^2 = \theta_3(1-\theta_3) + \omega_2^2$.

Define $\gamma_1 = \Pr\{T < c, D = 1\}$ and $\gamma_0 = \Pr\{T < c, D = 0\}$. Then, $\gamma_1 = \theta_1 - \theta_2$ and $\gamma_0 = 1 - \theta_1 - \theta_3$. Let

$$\hat{\gamma}_1 = \frac{1}{n}\sum_{i=1}^{n} I(T_i < c)\{V_i D_i + (1-V_i)\hat{\rho}_{Ki}\}$$

and

$$\hat{\gamma}_0 = \frac{1}{n}\sum_{i=1}^{n} I(T_i < c)\{V_i(1-D_i) + (1-V_i)(1-\hat{\rho}_{Ki})\}$$

be the KNN imputation estimators of γ_1 and γ_0, respectively. Let ζ_1^2 and ζ_0^2 denote the asymptotic variances of $\hat{\gamma}_1$ and $\hat{\gamma}_0$, respectively. The above given arguments still hold, leading to the expressions $\zeta_1^2 = \gamma_1(1-\gamma_1) + \omega_3^2$ and $\zeta_0^2 = \gamma_0(1-\gamma_0) + \omega_3^2$, where

$$\omega_3^2 = E[I(T < c)\rho(Y)(1-\rho(Y))(1-\pi(Y))]\left(1+\frac{1}{K}\right) + E\left[\frac{I(T < c)\rho(Y)(1-\rho(Y))(1-\pi(Y))^2}{\pi(Y)}\right].$$

It is easy to see that $\hat{\gamma}_1 = \hat{\theta}_1 - \hat{\theta}_2$ and $\hat{\gamma}_0 = 1 - \hat{\theta}_1 - \hat{\theta}_3$. As a consequence, the asymptotic covariances between $\hat{\theta}_1$ and $\hat{\theta}_2$ – say σ_{12} – and $\hat{\theta}_1$ and $\hat{\theta}_3$ – say σ_{13} – may be obtained as $\sigma_{12} = (1/2)(\sigma_1^2 + \sigma_2^2 - \zeta_1^2)$ and $\sigma_{13} = (1/2)(\sigma_1^2 + \sigma_3^2 - \zeta_0^2)$.

Finally, recall that $\widehat{Se}_{KNN}(c) = \frac{\hat{\theta}_2}{\hat{\theta}_1}$ and $\widehat{Sp}_{KNN}(c) = \frac{1-\hat{\theta}_1-\hat{\theta}_3}{1-\hat{\theta}_1}$. Therefore, by applying the delta method, one obtains

$$asVar(\widehat{Se}_{KNN}(c)) = \frac{\theta_2^2}{\theta_1^4}\sigma_1^2 + \frac{\sigma_2^2}{\theta_1^2} - 2\frac{\theta_2}{\theta_1^3}\sigma_{12}$$

and

$$asVar(\widehat{Sp}_{KNN}(c)) = \frac{\theta_3^2}{(1-\theta_1)^4}\sigma_1^2 + \frac{\sigma_3^2}{(1-\theta_1)^2} - 2\frac{\theta_3}{(1-\theta_1)^3}\sigma_{13}.$$

To obtain consistent estimates of the asymptotic variances given above, we may replace the unknown quantities in their expressions by the corresponding estimates. In particular, to estimate $asVar(\widehat{Se}_{KNN}(c))$ and $asVar(\widehat{Sp}_{KNN}(c))$, we ultimately need the estimates $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, $\hat{\omega}_1^2$, $\hat{\omega}_2^2$ and $\hat{\omega}_3^2$.

In a nonparametric regression imputation framework, quantities as ω_1^2, ω_2^2 and ω_3^2 are typically estimated by their empirical counterparts. The propensity score $\pi(y)$ is generally estimated by some kernel regression method (see Cheng [12]). In our context, however, we propose an approach that uses a nearest-neighbor rule to estimate both the functions $\rho(y)$ and $\pi(y)$ in ω_1^2, ω_2^2 and ω_3^2. In particular, for the conditional probabilities of disease we can use the estimates $\tilde{\rho}_i = \hat{\rho}_{\bar{K}i}$, for some suitable positive integer $\bar{K}$. For the conditional probabilities of verification, instead, we choose the estimates $\tilde{\pi}_i = \frac{1}{K_i^*}\sum_{j=1}^{K_i^*} V_{i(j)}$, where $\{(Y_{i(j)}, V_{i(j)}), j = 1, \ldots, K_i^*\}$ is a set of K_i^* observed pairs, and $Y_{i(j)}$ denotes the jth nearest neighbor to $Y_i = (T_i, X_i)^\top$ among all Y's. When $\tilde{\pi}_i$ is computed for a non-verified sample unit i, K_i^* is set equal to the rank of the first verified nearest neighbor to the unit i, i.e., K_i^* is such that $V_i = V_{i(1)} = V_{i(2)} = \cdots V_{i(K_i^*-1)} = 0$, and $V_{i(K_i^*)} = 1$. When $\tilde{\pi}_i$ is computed for a verified sample unit i, K_i^* is set equal to the rank of the first non-verified nearest neighbor to the unit i, i.e., K_i^* is such that $V_i = V_{i(1)} = V_{i(2)} = \cdots V_{i(K_i^*-1)} = 1$, and $V_{i(K_i^*)} = 0$. Observe that such procedure automatically avoids zero values for the $\tilde{\pi}_i$'s. Then, based on the $\tilde{\rho}_i$'s and the $\tilde{\pi}_i$'s, we obtain the estimates

$$\hat{\omega}_1 = \frac{K+1}{nK}\sum_{i=1}^{n}\tilde{\rho}_i(1-\tilde{\rho}_i)(1-\tilde{\pi}_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{\tilde{\rho}_i(1-\tilde{\rho}_i)(1-\tilde{\pi}_i)^2}{\tilde{\pi}_i},$$

$$\hat{\omega}_2 = \frac{K+1}{nK}\sum_{i=1}^{n} I(T_i \geq c)\tilde{\rho}_i(1-\tilde{\rho}_i)(1-\tilde{\pi}_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{I(T_i \geq c)\tilde{\rho}_i(1-\tilde{\rho}_i)(1-\tilde{\pi}_i)^2}{\tilde{\pi}_i}$$

and

$$\hat{\omega}_3 = \frac{K+1}{nK}\sum_{i=1}^{n} I(T_i < c)\tilde{\rho}_i(1-\tilde{\rho}_i)(1-\tilde{\pi}_i) + \frac{1}{n}\sum_{i=1}^{n}\frac{I(T_i < c)\tilde{\rho}_i(1-\tilde{\rho}_i)(1-\tilde{\pi}_i)^2}{\tilde{\pi}_i},$$

from which, together with $\hat{\theta}_1$, $\hat{\theta}_2$, $\hat{\theta}_3$, one derives the estimates of the variances of the KNN imputation estimators proposed in the paper. Clearly, to avoid $\hat{\omega}_1$, $\hat{\omega}_2$ and $\hat{\omega}_3$ to be equal to zero, we need to choose $\bar{K} > 1$ in estimating the conditional probabilities of disease.

To assess the behavior of the discussed variance estimators, we performed some simulation experiments. The results are given in Table 5. For the parameters θ_1, θ_2, θ_3, γ_1, γ_0, $Se(c)$ and $Sp(c)$, the table shows the relative biases, computed as $(MCV - MCM)/MCV$, where MCV is the Monte Carlo variance (multiplied by the sample size n, so as to obtain the asymptotic Monte Carlo variance) of the 1NN and 3NN estimators and MCM the Monte Carlo mean of the corresponding estimators of the asymptotic variances. The considered variance estimators are those discussed in this section. For each variance estimator, the involved estimates of parameters such as θ_1, θ_2, θ_3 are based on the same nearest-neighbor rule (1NN or 3NN) used to estimate the parameter of interest. For the estimates of the probabilities of disease in ω_1^2, ω_2^2 and ω_3^2, we chose $\bar{K} = 2$. The simulation setting is the same as in scenario (i) in Section 4, with some values for the pair (α, c) and sample size $n = 100$. The number of replicates in each simulation experiment is 5,000. Some other simulation results referring to scenario (ii) can be found in Section S3, Supplementary Material.

In summary, results in Table 5 (and in Section S3, Supplementary Material) seem to indicate that the proposed variance estimators behave satisfactorily. Of course, other variance estimators could be retrieved. For example one could, at least in principle, resort on resampling strategies. Naturally, this requires further investigation.

7 Discussion

This paper considers the estimation of the ROC curve of a continuous test under verification bias. Existing methods for correcting verification bias require estimation of $\rho(y)$ or $\pi(y)$, or both, and parametric models are commonly used to this end. However, as shown also by the simulation results presented in Section 4, a

Table 5: For each parameter: relative biases, computed as $(MCV - MCM)/MCV$, of the estimators of the asymptotic variance of KNN estimators.

			θ_1	θ_2	θ_3	γ_1	γ_0	$Se(c)$	$Sp(c)$
$\alpha = 0.5$	$c = 0.2$	1NN	0.092	0.027	−0.011	0.126	0.049	0.125	−0.011
		3NN	0.081	0.000	−0.030	0.077	0.036	0.067	−0.034
	$c = 0.5$	1NN	0.075	−0.038	0.030	0.131	0.069	0.071	0.036
		3NN	0.066	−0.055	0.000	0.104	0.053	0.046	0.006
	$c = 0.8$	1NN	0.126	0.022	−0.026	0.134	0.125	0.067	−0.030
		3NN	0.118	0.000	−0.086	0.112	0.116	0.047	−0.063
$\alpha = 1$	$c = 0.2$	1NN	0.048	0.024	0.027	0.087	0.034	0.098	0.031
		3NN	0.045	0.020	0.006	0.000	0.027	0.022	0.011
	$c = 0.5$	1NN	0.022	−0.011	−0.056	0.060	0.024	0.082	−0.050
		3NN	0.032	−0.017	−0.078	0.022	0.020	0.040	−0.070
	$c = 0.8$	1NN	0.022	−0.013	−0.017	0.067	0.004	0.073	−0.020
		3NN	0.023	−0.019	−0.056	0.048	0.004	0.049	−0.043
$\alpha = 1.5$	$c = 0.2$	1NN	0.023	0.014	0.022	−0.250	0.016	−0.067	0.021
		3NN	0.028	0.015	0.006	−0.250	0.020	−0.170	0.007
	$c = 0.5$	1NN	0.023	0.025	−0.008	0.000	−0.013	0.031	−0.014
		3NN	0.024	0.020	−0.016	−0.083	−0.013	−0.034	−0.024
	$c = 0.8$	1NN	0.050	0.021	−0.012	0.000	0.021	0.014	−0.014
		3NN	0.047	0.016	−0.025	−0.036	0.017	−0.010	−0.030

Notes: MCV is the Monte Carlo variance (multiplied by the sample size n) of the 1NN and 3NN estimators and MCM is the Monte Carlo mean of the corresponding estimators of the asymptotic variances. The considered variance estimators are those discussed in Section 6. For each variance estimator, the estimates of parameters such as θ_1, θ_2, θ_3 are based on the same nearest-neighbor rule (1NN or 3NN) used to estimate the parameter of interest. For the estimates of the probabilities of disease in ω_1^2, ω_2^2 and ω_3^2, it is $\bar{K} = 2$. The simulation setting is the same as in scenario (i) in Section 4, with some values for the pair (α, c). The sample size is $n = 100$. The number of replicates in each simulation experiment is 5,000.

wrong specification of these models can have an adverse impact on the performance of the estimators, which result in a high bias and/or unstable behavior.

The new estimators of sensitivity and specificity (6) are fully nonparametric. Their use reduces the effects of possible misspecification to the inference results. The loss of efficiency with respect to the use of parametric competitors (when these can be reasonably employed) can range from minimal to sensible values according to the nature of the problem at hand, as simulation results in the main paper and in Supplementary Material, Section S4 show. This is somehow intrinsic in the nonparametric nature of the proposed estimators.

The new approach is based on the K-nearest-neighbor imputation, which requires the choice of a value for K. Our simulation results (see also the Supplementary Material) seem to confirm results in Ning and Cheng [9] according to which a small value of K-within the range 1–3 may be a good choice. It is worth noting, however, that the choice of K might depend upon the dimension of the feature space. In our study, the feature space includes the diagnostic test result T and the an auxiliary covariate X of dimension 1 (and 3, see Section S4, Supplementary Material). A small number of features is quite common in the context of the evaluation of diagnostic tests. However, if the number of features increases, it could be convenient to consider higher values for K. Of course, the nonparametric nature of the approach imposes to take into account the number of verified units, both in the healthy and diseased group, available in the sample. In particular, this means that K should not be too big compared to the number of the verified subjects, n_{ver} say. Generally speaking, a possible strategy to choose a suitable value for K in practice could be cross-validation, based on using the KNN estimators on the verified subjects only. Each verified subject is treated in turn as if it were not verified; for fixed K, the estimate $\hat{\rho}_{Ki}$ of its conditional disease probability is computed using KNN imputation and compared to the truth to produce a measure of discrepancy. This is done for a number of K's and the K for which the discrepancy is smallest is retained for use in the original sample. A possible choice for the discrepancy in this context could be $\frac{1}{n_{ver}} \sum_{i=1}^{n_{ver}} |D_i - \hat{\rho}_{Ki}|$.

The issue of the choice of the distance measure to use is of more general nature. Our simulation results (see Supplementary Material) seem to indicate that the standard Euclidean distance may be a good choice. However, it is clear that an adequate choice ultimately depends on several aspects, such as features of the data to analyze, as well as computational concerns.

Estimators (6) modify in an obvious way when no covariates are measured, i.e., when $Y = T$. Moreover, a simple extension, that could be used when categorical variables are also observed for each patient, is possible. Consider, for example, the problem of estimating the sensitivity. Without loss of generality, suppose that a single factor U, with u levels, is observed together with Y. We also assume that U may be associated with both D and V. Then, if Theorem 1 holds in each stratum, i.e., in each group of units with the same level of U, a consistent and asymptotically normally distributed estimator of $Se_{KNN}(c)$ is

$$\frac{1}{n}\sum_{j=1}^{u} \widehat{Se}^{cond}_{KNN_j}(c) n_j,$$

where n_j denotes the size of the jth sample stratum and $\widehat{Se}^{cond}_{KNN_j}(c)$ is the KNN estimator of the conditional sensitivity, i.e., $\widehat{Se}_{KNN}(c)$ obtained from the patients in the jth stratum. Clearly, the use of such estimator relies on availability of sufficient information in each stratum.

As suggested by a Referee, one could think of possible devices aimed at enhancing performances of the estimators. One possibility could be to assign unequal weights to the K nearest neighbors entering in the estimates $\hat{\rho}_{Ki}$. This might produce a reduction of the mean square error of the KNN estimators for the sensitivity and the specificity. Otherwise, "hybrid" estimators for the sensitivity and specificity could be obtained by combining the SPE estimator with the KNN strategy. This could lead, at least in principle, to partially parametric estimators, robust with respect to possible weaknesses of the (nonparametric) model chosen for $\rho(y)$ when, for example, the disease process depends also on unobserved auxiliary variables. These are interesting and intriguing topics whose development, however, requires non-trivial treatment, both from a theoretical and empirical perspective.

Acknowledgement: The contribution of Stefano Mussi in producing some simulation results in Supplementary Material is gratefully acknowledged.

References

1. Begg CB, Greenes RA. Assessment of diagnostic tests when disease verification is subject to selection bias. Biometrics 1983;39:207–15.
2. Begg CB. Biases in the assessment of diagnostic tests. Statist Med 1987;6:411–23.
3. Zhou X-H. Correcting for verification bias in studies of a diagnostic test's accuracy. Stat Meth Med Res 1998;7:337–53.
4. Zhou X-H, Obuchowski NA, McClish DK. Statistical methods in diagnostic medicine. New York: Wiley-Interscience, 2002.
5. Alonzo TA, Pepe MS. Assessing accuracy of a continuous screening test in the presence of verification bias. J R Stat Soc Ser C 2005;54:173–90.
6. Fluss R, Reiser B, Faraggi D, Rotnitzky A. Estimation of the ROC curve under verification bias. Biometrical J 2009;51:475–90.
7. Liu D, Zhou XH. A model for adjusting for nonignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach. Biometrics 2010;66:1119–28.
8. He H, McDermott MP. A robust method using propensity score stratification for correcting verification bias for binary tests. Biostatistics 2012;13:32–47.
9. Ning J, Cheng PE. A comparison study of nonparametric imputation methods. Statist Comput 2012;22:273–85.
10. Alonzo TA, Pepe MS. Estimating disease prevalence in two-phase studies. Biostatistics 2003;4:313–23.
11. Asuncion A, Newman DJ. UCI machine learning repository. Irvine, CA: School of Information and Computer Science, University of California, 2007. Available at: http://www.ics.uci.edu/mlearn/MLRepository.html
12. Cheng PE. Nonparametric estimation of mean functionals with data missing at random. J Am Stat Assoc 1994;89:81–7.

Supplemental Material: The online version of this article (DOI: 10.1515/ijb-2014-0014) offers supplementary material, available to authorized users.