Statistics and Its Interface Volume 10 (2017) 629–641

Nonparametric verification bias-corrected inference for the area under the ROC curve of a continuous-scale diagnostic test

Gianfranco Adimari* and Monica Chiogna

For a continuous-scale diagnostic test, the area under the receiver operating characteristic curve (AUC) is a popular summary measure to assess the test's ability to discriminate between healthy and diseased subjects. In some studies, verification of the true disease status is performed only for a subset of subjects, selected possibly on the basis of the test result and of other characteristics of the subjects. Estimators of the AUC based only on this subset of subjects are typically biased; this is known as verification bias. Some methods have been proposed to correct verification bias, but they require parametric models for the (conditional) probability of disease and/or the (conditional) probability of verification. A wrong specification of such parametric models can affect the behaviour 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 AUC 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. The new AUC estimator is consistent under the assumption that the true disease status, if missing, is missing at random (MAR). A simple extension which deals with stratified samples is also provided. Simulation experiments are used to investigate the finite sample behaviour of the proposed methods. An illustrative example is presented.

AMS 2000 subject classifications: Primary 62G05, 62G20; secondary 62P10.
Keywords and phrases: Missing data imputation, Nearest-neighbor imputation, ROC analysis.

1. INTRODUCTION

The evaluation of the ability of a diagnostic test to separate diseased from non-diseased subjects is a crucial issue in modern medicine. 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 gold standard test, which assesses the 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's characteristics. However, summary measures of test performance based only on data from patients with verified disease status 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 test's ability to discriminate between healthy and diseased subjects. The continuous test result can be dichotomized at a specified cutpoint. Given a cutpoint, the sensitivity is the probability of a true positive, i.e., the probability that the test correctly identifies a diseased subject. The specificity 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 across the range of possible cutpoints is the area under the ROC curve (AUC), that can be interpreted as the probability that a randomly selected diseased case will have a test result worse (for example bigger, if large test values are more likely to be linked to disease) than a randomly selected nondiseased case. Reasonable values for the AUC range from 0.5 to 1. The larger the AUC value, the more accurate the diagnostic test is. An AUC of 0.5 means that the diagnostic accuracy in question is equivalent to that which would be obtained by flipping a coin (i.e., random chance).

In the presence of verification bias, under the assumption that the true disease status, if missing, is missing at random (MAR), estimation of the AUC of a continuous test is discussed in [5], where an estimator based on the inverse probability weighting approach is proposed. MAR assumption states 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. This corresponds to a so called ignorable missingness, which is often assumed in practice. Estimation of the AUC when the

*Corresponding author.

true diseased status is subject to non-ignorable missingness is tackled in [10] and [6]. Under different model settings, these papers develop AUC estimators based on imputation and/or reweighting methods. Clearly, such estimators apply also when the missingness is ignorable, i.e., under the stronger MAR assumption.

However, known methods to inference on the AUC of a continuous test require specification of parametric regression models for the probability of a subject being diseased and/or verified. Typically, suitable generalized linear regression models are employed to this end.

In real-world applications, a correct specification of the disease model and/or the verification model could be cumbersome. As suggested in [6], in some circumstances researchers could collect necessary information on the reason of missing gold standard and disease mechanism, and this makes it still possible to build approximately correct models from the scientific point of view. Without any information about the selection and disease mechanisms, a prudent approach that avoids misspecification problems consists in resorting to nonparametric methods. To the best of our knowledge, no such methods are available for inference on the AUC of a continuous-scale diagnostic test.

In this paper, we propose a fully nonparametric method for the estimation of the AUC under verification bias. The proposed method, which follows the lines drawn in [1] for the estimation of the ROC curve, 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 is verified. The estimator for the AUC obtained by the new approach is shown to be consistent under the MAR assumption. A simple extension which deals with stratified samples is also provided, and estimation of the standard deviation of the proposed estimators is discussed. Several simulation experiments are used to investigate the finite sample behaviour of our proposals. An application to a real dataset is also presented.

The paper is organized as follows. In Section 2, we describe the proposed approach, whose theoretical justification is given in Appendix 1, Supplementary Material (http://intlpress.com/site/pub/pages/journals/items/sii/content/vols/0010/0004/s002). Section 3 presents the results of four simulation studies, and Section 4 contains the illustrative example. Some final remarks are reported in Section 5.

2. THE PROPOSED METHOD

2.1 Nonparametric AUC estimator

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 that the i-th patient is diseased and $D_i = 0$ indicates that the i-th patient is free of disease. Without loss of generality, we assume that a high test result indicates a high probability of disease; then, the AUC represents the probability that a diseased subject has a larger test result than a non-diseased one, i.e., $AUC = \Pr(T_i > T_j | D_i = 1, D_j = 0)$.

Let V_i denote the binary verification status of the i-th patients, with $V_i = 1$ if the i-th patient has the true disease status verified, and $V_i = 0$ otherwise. In practice, some information, other than the results from the test, may be obtained for each patient. Let X_i be a vector of observed covariates for the i-th patient that may be associated with both D_i and V_i. Hereafter, we assume X to be a continuous-valued random vector.

When all patients are verified, i.e., $V_i = 1$, $i = 1, \ldots, n$, a complete data set is obtained. In this case, it is well known that a nonparametric AUC estimator is given by the Wilcoxon statistic

$$\frac{\sum_{i=1}^{n} \sum_{j=1,\ j\neq i}^{n} I(T_i > T_j) D_i (1 - D_j)}{\sum_{i=1}^{n} \sum_{j=1,\ j\neq i}^{n} D_i (1 - D_j)},$$

where $I(\cdot)$ is the indicator function. If only a subset of patients has the disease status verified, some labels D_i are missing in the sample (T_i, X_i, D_i, V_i), $i = 1, \ldots, n$. Hence, in order to obtain a verification bias-corrected version of the Wilcoxon statistic, one could use some suitable nonparametric imputation technique to impute the missing D_i's.

In what follows, we propose to use a K-nearest-neighbor (KNN) imputation method to obtain a nonparametric verification bias-corrected AUC estimator. Let $\rho_i = \Pr(D_i = 1 | T_i, X_i)$ denote the probability that the i-th patient is diseased given the test result and covariates. For a finite positive integer K and a suitable distance measure, a nearest-neighbor imputation estimate of ρ_i, for a subject with true disease status not verified, could be defined as

$$\hat{\rho}_{Ki} = \frac{1}{K} \sum_{j=1}^{K} D_{i(j)},$$

where $\{(Y_{i(j)}, D_{i(j)}) : V_{i(j)} = 1,\ j = 1, \ldots, K\}$ is a set of K observed data pairs and $Y_{i(j)}$ denotes the j-th nearest neighbor to $Y_i = (T_i, X_i^\top)^\top$ among all Y's corresponding to the verified patients, i.e., to those D_h's with $V_h = 1$. Then, the estimate $\hat{\rho}_{Ki}$ could be used as imputation value for the missing label D_i. This leads to the proposed nonparametric verification bias-corrected AUC estimator:

$$\widehat{AUC} = \frac{\sum_{i=1}^{n} \sum_{j=1,\ j\neq i}^{n} I(T_i > T_j) \tilde{D}_i (1 - \tilde{D}_j)}{\sum_{i=1}^{n} \sum_{j=1,\ j\neq i}^{n} \tilde{D}_i (1 - \tilde{D}_j)}, \tag{1}$$

with $\tilde{D}_i = V_i D_i + (1 - V_i)\hat{\rho}_{Ki}$. In Appendix 1, Supplementary Material, we show that, if the MAR assumption holds, i.e. $\Pr(V = 1 | D, Y) = \Pr(V = 1 | Y)$ where $Y = (T, X^\top)^\top$, the functions $\rho(y) = \Pr(D = 1 | Y = y)$ and $\pi(y) = \Pr(V = 1 | Y = y)$ are first-order differentiable and $E(1/\pi(Y)) < \infty$, then the KNN imputation estimator $\widehat{AUC}$, based on the sample (T_i, X_i, D_i, V_i), $i = 1, \ldots, n$, is consistent. Moreover, in the same appendix, elements arise that lead us to conjecture that the KNN imputation estimator is asymptot-

ically normally distributed, a conjecture also supported by results of numerical studies in Section 3.

Being X in our theoretical setting a continuous-valued random vector, distance ties in the nearest neighbors identification occur with probability 0. Nevertheless, in real applications, continuous variables are often measured on a discrete scale, so that distance ties may occur. In this case, it is necessary to identify a suitable tie-breaking strategy. Among the solutions proposed in the literature, a simple approach consists in artificially enlarging the feature space by adding to X a continuous random component, generated independently of all other variables (see [4], Chapter 11). This is the approach that we employ in the application of Section 4.

Estimator (1) modifies in an obvious way when no covariates are measured, i.e., when $Y = T$. Moreover, a simple extension of our AUC estimator, that could be used when categorical variables are also observed for each patient, is possible. Without loss of generality, we suppose that a single factor C, with m levels, is observed together with Y. We also assume that C may be associated with both D and V. In this case, the sample can be divided into m strata, i.e. m groups of units sharing the same level of C. Then, if the MAR assumption and first-order differentiability of the functions $\rho(y)$ and $\pi(y)$ hold in each stratum, a consistent estimator of the AUC of the test T is

$$\widehat{AUC}^S = \frac{1}{n}\sum_{j=1}^{m} \widehat{AUC}_j^{cond} n_j, \tag{2}$$

where n_j denotes the size of the j-th stratum and $\widehat{AUC}_j^{cond}$ is the KNN estimator of the conditional AUC, i.e., the KNN AUC estimator (1) obtained from the patients in the j-th stratum. Of course, we must assume that, for every j, ratios n_j/n have finite and nonzero limits as n goes to infinity.

Finally, we observe that, from a theoretical point of view, the use of the proposed estimator is not restricted to any special type of distance measure, nor to any particular choice of the neighborhood size K. However, to apply the method, such choices have to be taken. These aspects will be discussed in the Section 3, where the impact of the choices of K and the distance measure on the estimator's performance will be empirically investigated in a simulation scenario.

2.2 Confidence intervals

An important issue, which is crucial in many applications, is the estimation of the standard deviation of the proposed KNN AUC estimator. To address the problem, we propose to employ a simple bootstrap procedure, based on the following steps. From the sample at hand (T_i, X_i, D_i, V_i), $i = 1, \ldots, n$, obtain B bootstrap samples $(T_i^{*b}, X_i^{*b}, D_i^{*b}, V_i^{*b})$, $b = 1, \ldots, \mathrm{B}$, and $i = 1, \ldots, n$, and compute the bootstrap estimates $\widehat{AUC}^{*b}$. Due to the nature of bootstrap samples, each bootstrap estimate needs to be computed using a version of (1) that takes into account the presence of ties in the sample, i.e.,

$$\widehat{AUC}^{*b} = \frac{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} I_{ij}^{*b}\tilde{D}_i^{*b}(1-\tilde{D}_j^{*b})}{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} \tilde{D}_i^{*b}(1-\tilde{D}_j^{*b})}, \tag{3}$$

with $I_{ij}^{*b} = I(T_i^{*b} > T_j^{*b}) + 0.5I(T_i^{*b} = T_j^{*b})$, and $\tilde{D}_i^{*b} = V_i^{*b}D_i^{*b} + (1 - V_i^{*b})\hat{\rho}_{Ki}^{*b}$. Here, $\hat{\rho}_{Ki}^{*b}$ denotes the nearest-neighbor imputation value for a missing label D_i^{*b} in the bootstrap sample. Then, the bootstrap estimator of the standard deviation of $\widehat{AUC}$ is

$$\widehat{sd} = \sqrt{\frac{1}{\mathrm{B}-1}\sum_{b=1}^{\mathrm{B}}\left(\widehat{AUC}^{*b} - \widehat{AUC}^{*}\right)^2},$$

where $\widehat{AUC}^{*}$ is the mean of the B bootstrap estimates $\widehat{AUC}^{*b}$.

Once the estimate of the standard deviation of the KNN estimator is obtained, under the conjectured asymptotic normality of the estimator, inference on the AUC can be made using the pivot $(\widehat{AUC} - AUC)/\widehat{sd}$. In particular, a confidence interval with nominal coverage $1 - \gamma$ is given by $\widehat{AUC} \pm z_{1-\gamma/2}\widehat{sd}$, where z_γ denotes the γ-quantile of a standard normal random variable. It is known that confidence intervals based on the crude normal approximation approach may behave poorly when the sample size is small. A classical method to improve their accuracy is based on appropriately transforming the parameter of interest. Suitable transformations, which turn out to be useful in this setting are, for instance, the logit and the probit transformation. In particular, considering the logit transformation, one can set $\zeta = \ell(AUC) = \log\left(\frac{AUC}{1-AUC}\right)$, $\hat{\zeta} = \ell(\widehat{AUC})$, so that $(\hat{\zeta} - \zeta)$ is approximately normal with mean zero and standard deviation estimated by $\frac{\widehat{sd}}{\widehat{AUC}(1-\widehat{AUC})}$. This result can be used to construct confidence intervals on the ζ scale, which are then converted back to the AUC scale by the inverse transformation ℓ^{-1}. The transformation also provides range-respecting confidence intervals.

Both effectiveness of the suggested bootstrap procedure and accuracy of the above discussed confidence intervals based on $\widehat{AUC}$ will be investigated by the simulation studies described in the next section. Of course, there are other methods to obtain confidence intervals via bootstrap procedures. For instance, one could resort to bootstrap calibration to retrieve the appropriate quantiles of the distribution of the estimator $\widehat{AUC}$. This method does not depend on the normal approximation of the estimator; but is computationally more expensive, and in our view, less manageable for practitioners.

3. SIMULATION STUDIES

In this section, we present the results of four Monte Carlo studies. In the first study, the aim is to assess the behaviour of the proposed KNN AUC estimator (1) in samples of small to moderate sample sizes, also investigating the effects of the

choice of K as well as of the distance measure for the definition of the neighborhood. In the second study, the objective is to evaluate the performance of the bootstrap estimator for the standard deviation of the proposed estimator. In the third study, we assess the behaviour of estimator (2) which deals with stratified samples and finally, in the last study, we compare our proposal with alternative estimators. Such estimators, discussed in [10], [5], and [6], require specification of parametric regression models for the disease and/or the verification processes. For the reader's convenience, we report here the expression of such estimators. The full imputation (FI) estimator is

$$\widehat{AUC}_{FI} = \frac{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} I(T_i > T_j)\hat{\rho}_i(1-\hat{\rho}_j)}{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} \hat{\rho}_i(1-\hat{\rho}_j)}.$$

Parametric models, such as logistic regression models, have to be specified to obtain the estimates $\hat{\rho}_i$ of $\rho_i = \Pr(D_i = 1|T_i, X_i)$, $i = 1, \ldots, n$, using only data from verified subjects. Mean score imputation (MSI) is another possible approach that only imputes the disease status for unverified subjects. In this case,

$$\widehat{AUC}_{MSI} = \frac{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} I(T_i > T_j)\hat{D}_i(1-\hat{D}_j)}{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} \hat{D}_i(1-\hat{D}_j)},$$

with $\hat{D}_i = V_i D_i + (1-V_i)\hat{\rho}_i$. 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 estimator is

$$\widehat{AUC}_{IPW} = \frac{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} I(T_i > T_j)V_i D_i V_j(1-D_j)/(\hat{\pi}_i\hat{\pi}_j)}{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} V_i D_i V_j(1-D_j)/(\hat{\pi}_i\hat{\pi}_j)},$$

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

$$\widehat{AUC}_{SPE} = \frac{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} I(T_i > T_j)\hat{D}_i(1-\hat{D}_j)}{\sum_{i=1}^{n}\sum_{j=1,\ j\neq i}^{n} \hat{D}_i(1-\hat{D}_j)},$$

with $\hat{D}_i = V_i D_i/\hat{\pi}_i + (1-V_i/\hat{\pi}_i)\hat{\rho}_i$. Under MAR assumption, the SPE estimator is doubly robust in the sense that it is consistent if either π_i's or ρ_i's are estimated consistently.

Study 1 and Study 2. Simulation settings in the first two studies are similar to those in [2] and [5]. 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 $D = I(g(Z_1, Z_2) > \nu)$. The threshold ν determines the disease prevalence 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, $T = h(Z_1, Z_2) + \epsilon_1$ and $X = f(Z_1, Z_2) + \epsilon_2$, 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. The verification probability π is set to be a suitable function of T and X, in accordance with the MAR assumption. In particular, we set $g(Z_1, Z_2) = Z_1 + Z_2$, $f(Z_1, Z_2) = \beta(Z_1 + Z_2)$, $h(Z_1, Z_2) = \alpha(Z_1 + Z_2)$, and $\pi(T, X) = \frac{e^{\delta_0+\delta_1 T+\delta_2 X}}{1+e^{\delta_0+\delta_1 T+\delta_2 X}}$. 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. Moreover, we choose ν to make the disease prevalence equal to 0.25. As for α, we choose five different values, i.e, 0.1, 0.25, 0.5, 1 and 1.5, giving rise to five different values for the true AUC, i.e., 0.595, 0.714, 0.846, 0.943, 0.973, respectively. Finally, we set $\beta = 1$ or $\beta = 0.1$: in the first case, the resulting covariate X has itself high accuracy (AUC of 0.943), while, in the latter case, its accuracy is low (AUC approximately equal to 0.6). The number of replications in each simulation experiment is 5000.

In the **first study**, we fix three sample sizes ($n = 75, 100$, and 200) and consider different distance measures to define the estimators. Among the various possibilities, we consider here the most commonly used distances, i.e., the Euclidean distance, the Manhattan distance, the Lagrange distance and the Mahalanobis distance (see Appendix 2, Supplementary Material, for the definitions). Table 1 shows Monte Carlo means and standard deviations of the new KNN estimators, with $K = 1$, $K = 3$, $K = 5$, $K = 10$ and $K = 20$, based on the Euclidean distance. Each block in the table refers to a chosen pair (β, sample size). Rows denoted by "Full" indicate the results for the Wilcoxon statistic, i.e., the AUC estimator based on complete data, which is used as benchmark. Complete results of the study covering all considered distance measures are given in Appendix 3, Supplementary Material. Performances of the KNN estimators are quite comparable for different choices of the distance measure, with the Mahalanobis distance performing slightly worse than competing distances. This might be due to the fact that, in our simulation settings, there is not a large disparity in the range of the data in each dimension. Results show also that bias increases on increasing the number of nearest neighbors – although such an effect tends to attenuate for increasing sample sizes – indicating that the use of values of K which are large compared to the number of the verified units in the sample fails to represent the local pattern of the measurement space, i.e., of the Y space. This might also be related to the dimension of the Y space. In this study, where the feature space includes the diagnostic test result T and the unidimensional auxiliary covariate X, the evidence is that the choice of a small value of K (within the range 1 to 3) seems a good choice. Similar conclusions also come from the results of Study 4, where a covariate X of dimension three is employed. Finally, from the results in Table 1 (and Tables 1–6 in Appendix 3, Supplementary Material) we observe that the standard deviation of the estimators increases when both α and β are small, i.e. when both X and T are poorly associated with D.

In the **second study**, we consider the bootstrap estimator for the standard deviation of our KNN estimators

Table 1. Study 1. Monte Carlo means and standard deviations of the KNN AUC estimators, for different values of α, different choices of K and the Euclidean distance measure

	$\alpha = 0.1$		$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 1.5$	
	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.
					$\beta = 1$ Sample size = 75					
Full	0.594	0.077	0.717	0.069	0.845	0.050	0.944	0.027	0.973	0.016
1NN	0.594	0.094	0.717	0.086	0.845	0.064	0.944	0.034	0.973	0.021
3NN	0.594	0.089	0.715	0.080	0.843	0.060	0.942	0.033	0.972	0.020
5NN	0.593	0.086	0.714	0.078	0.841	0.059	0.940	0.033	0.970	0.021
10NN	0.590	0.082	0.708	0.075	0.832	0.059	0.933	0.035	0.964	0.023
20NN	0.579	0.073	0.683	0.070	0.793	0.063	0.887	0.053	0.920	0.050
					$\beta = 0.1$ Sample size = 75					
Full	0.592	0.078	0.717	0.069	0.846	0.050	0.943	0.027	0.973	0.017
1NN	0.591	0.122	0.713	0.109	0.841	0.081	0.942	0.041	0.973	0.024
3NN	0.587	0.110	0.705	0.099	0.834	0.075	0.938	0.039	0.970	0.023
5NN	0.584	0.106	0.699	0.095	0.827	0.075	0.933	0.039	0.966	0.024
10NN	0.577	0.097	0.686	0.089	0.810	0.074	0.920	0.042	0.955	0.026
20NN	0.567	0.081	0.662	0.077	0.772	0.069	0.879	0.050	0.917	0.040
					$\beta = 1$ Sample size = 100					
Full	0.595	0.066	0.716	0.060	0.846	0.043	0.944	0.023	0.972	0.014
1NN	0.595	0.080	0.716	0.072	0.847	0.054	0.944	0.030	0.973	0.018
3NN	0.595	0.076	0.715	0.068	0.846	0.050	0.943	0.028	0.972	0.017
5NN	0.594	0.075	0.714	0.067	0.844	0.050	0.941	0.028	0.971	0.017
10NN	0.592	0.072	0.710	0.065	0.839	0.049	0.937	0.028	0.967	0.018
20NN	0.587	0.067	0.697	0.062	0.821	0.051	0.918	0.036	0.952	0.026
					$\beta = 0.1$ Sample size = 100					
Full	0.595	0.067	0.717	0.059	0.847	0.043	0.943	0.023	0.973	0.014
1NN	0.590	0.104	0.712	0.093	0.844	0.067	0.943	0.035	0.972	0.021
3NN	0.588	0.094	0.707	0.084	0.837	0.063	0.940	0.033	0.970	0.020
5NN	0.585	0.091	0.702	0.082	0.832	0.062	0.936	0.034	0.968	0.020
10NN	0.581	0.085	0.692	0.078	0.819	0.062	0.927	0.035	0.961	0.022
20NN	0.573	0.075	0.673	0.070	0.793	0.061	0.902	0.039	0.939	0.026
					$\beta = 1$ Sample size = 200					
Full	0.595	0.047	0.714	0.041	0.846	0.030	0.943	0.016	0.973	0.010
1NN	0.595	0.057	0.714	0.051	0.846	0.038	0.944	0.020	0.973	0.013
3NN	0.595	0.054	0.714	0.048	0.845	0.036	0.943	0.019	0.972	0.012
5NN	0.595	0.054	0.713	0.047	0.845	0.036	0.942	0.019	0.972	0.012
10NN	0.594	0.052	0.712	0.046	0.843	0.035	0.940	0.019	0.970	0.012
20NN	0.592	0.051	0.709	0.045	0.838	0.035	0.936	0.019	0.967	0.012
					$\beta = 0.1$ Sample size = 200					
Full	0.594	0.046	0.716	0.041	0.846	0.031	0.944	0.017	0.972	0.010
1NN	0.592	0.075	0.713	0.066	0.844	0.047	0.943	0.024	0.972	0.015
3NN	0.589	0.068	0.710	0.060	0.840	0.044	0.941	0.023	0.971	0.014
5NN	0.588	0.067	0.707	0.059	0.837	0.044	0.940	0.023	0.970	0.014
10NN	0.585	0.064	0.701	0.057	0.830	0.044	0.935	0.023	0.968	0.014
20NN	0.580	0.060	0.690	0.055	0.817	0.044	0.926	0.024	0.961	0.015

described in Section 2.2. Given the simulation results obtained in the first study, we focus our attention on 1NN and 3NN estimators based on the Euclidean distance measure. For each Monte Carlo sample of size n, we compute the AUC estimates and the bootstrap standard deviations based on 200 bootstrap replications.

Table 2 and Table 3 contain the Monte Carlo means of the bootstrap estimates of the standard deviations and the Monte Carlo standard deviations for the KNN AUC estimators at the chosen values of α and for different sample sizes, for $\beta = 1$ and $\beta = 0.1$, respectively. Sample sizes increase with increasing values of α, i.e., of the true AUC values. In particular, we fix 100 and 200 when α is 0.1; 150 and 300 when α is 0.25; 200 and 400 when α is 0.5; 250 and 500 when α is 1; 300 and 600 when α is 1.5. Table 2 and Table 3 also report the empirical coverages of the confidence intervals

Table 2. Study 2. Monte Carlo means of bootstrap standard deviations (bootstrap s.d.) and Monte Carlo standard deviations (MC s.d.) for the 1NN and 3NN AUC estimators based on the Euclidean distance measure, for $\beta = 1$, different values of α and different sample sizes. In the right side, empirical coverages of the confidence intervals for the AUC obtained through the normal approximation approach, with the standard deviation of $\widehat{AUC}$ estimated using the bootstrap method. Nominal coverages: 0.99, 0.95, 0.90. Figures in bold denote the empirical coverages raised by the normal approximation method after logit transormation

α	sample size		bootstrap s.d.	MC s.d.	empirical coverages		
0.1	100	1NN	0.078	0.080	0.980	0.934	0.879
					0.989	**0.949**	**0.892**
		3NN	0.077	0.076	0.982	0.943	0.892
					0.990	**0.957**	**0.907**
	200	1NN	0.055	0.056	0.981	0.933	0.881
		3NN	0.054	0.054	0.986	0.941	0.893
0.25	150	1NN	0.057	0.058	0.979	0.926	0.875
					0.992	**0.947**	**0.891**
		3NN	0.056	0.055	0.981	0.936	0.890
					0.993	**0.955**	**0.901**
	300	1NN	0.040	0.041	0.982	0.934	0.880
		3NN	0.039	0.039	0.987	0.944	0.898
0.5	200	1NN	0.036	0.038	0.969	0.919	0.866
					0.988	**0.942**	**0.882**
		3NN	0.036	0.036	0.975	0.932	0.884
					0.988	**0.950**	**0.895**
	400	1NN	0.026	0.027	0.978	0.930	0.877
		3NN	0.025	0.025	0.983	0.940	0.890
1	250	1NN	0.017	0.018	0.958	0.909	0.859
					0.990	**0.941**	**0.880**
		3NN	0.017	0.017	0.968	0.924	0.875
					0.992	**0.949**	**0.897**
	500	1NN	0.012	0.013	0.972	0.919	0.867
		3NN	0.012	0.012	0.976	0.930	0.877
1.5	300	1NN	0.010	0.010	0.956	0.913	0.865
					0.987	**0.941**	**0.885**
		3NN	0.090	0.090	0.964	0.923	0.879
					0.990	**0.950**	**0.899**
	600	1NN	0.007	0.007	0.971	0.924	0.872
		3NN	0.007	0.007	0.978	0.936	0.888

for the AUC obtained through the normal approximation approach, with the standard deviation of $\widehat{AUC}$ estimated using the bootstrap method. The considered confidence intervals have nominal coverage 0.99, 0.95, 0.90. For each value of α and the corresponding smallest sample size, the tables also give the empirical coverages raised by the normal approximation approach after logit transformation.

Results given in the tables seem to show effectiveness of the bootstrap procedure in estimating the standard deviation of the KNN AUC estimators and its usefulness in the construction of confidence intervals. Evidently, also the conjecture about asymptotic normality of $\widehat{AUC}$ seems to be supported. Overall, the 3NN AUC estimator seems to achieve better results than the 1NN AUC estimator when the objective of inference is the construction of confidence intervals. Clearly, in practical situations, as suggested also by the simulation results, we expect that the sample size needed to achieve sufficiently accurate confidence intervals

Table 3. Study 2. Monte Carlo means of bootstrap standard deviations (bootstrap s.d.) and Monte Carlo standard deviations (MC s.d.) for the 1NN and 3NN AUC estimators based on the Euclidean distance measure, for $\beta = 0.1$, different values of α and different sample sizes. In the right side, empirical coverages of the confidence intervals for the AUC obtained through the normal approximation approach, with the standard deviation of $\widehat{AUC}$ estimated using the bootstrap method. Nominal coverages: 0.99, 0.95, 0.90. Figures in bold denote the empirical coverages raised by the normal approximation method after logit transormation

α	sample size		bootstrap s.d.	MC s.d.	empirical coverages		
0.1	100	1NN	0.104	0.099	0.965	0.921	0.869
					0.988	**0.944**	**0.889**
		3NN	0.094	0.096	0.974	0.933	0.889
					0.989	**0.951**	**0.909**
	200	1NN	0.073	0.069	0.974	0.918	0.863
		3NN	0.066	0.067	0.982	0.938	0.894
0.25	150	1NN	0.076	0.071	0.968	0.915	0.855
					0.986	**0.935**	**0.879**
		3NN	0.070	0.069	0.975	0.932	0.881
					0.988	**0.945**	**0.893**
	300	1NN	0.052	0.049	0.977	0.927	0.873
		3NN	0.048	0.049	0.985	0.942	0.889
0.5	200	1NN	0.047	0.044	0.963	0.916	0.868
					0.984	**0.935**	**0.885**
		3NN	0.044	0.044	0.975	0.938	0.889
					0.988	**0.946**	**0.899**
	400	1NN	0.033	0.031	0.975	0.921	0.871
		3NN	0.031	0.031	0.981	0.942	0.896
1	250	1NN	0.021	0.020	0.948	0.899	0.844
					0.983	**0.935**	**0.879**
		3NN	0.020	0.020	0.961	0.914	0.877
					0.984	**0.943**	**0.892**
	500	1NN	0.015	0.014	0.965	0.914	0.861
		3NN	0.014	0.014	0.973	0.933	0.886
1.5	300	1NN	0.011	0.010	0.947	0.897	0.848
					0.982	**0.935**	**0.875**
		3NN	0.011	0.010	0.957	0.917	0.875
					0.985	**0.944**	**0.889**
	600	1NN	0.008	0.008	0.965	0.920	0.868
		3NN	0.007	0.007	0.975	0.938	0.892

depends on the true AUC value and on the rate of verified units (healthy as well as diseased) in the sample. High values of AUC and small verification rates will likely require a high number of sample units. Generally speaking, the logit transformation seems to greatly improve the coverage accuracy.

Study 3. We consider a simulation setting similar to that adopted in studies 1 and 2, with $Z_1 \sim N(0, 0.5)$, $Z_2 \sim N(0, 0.5)$, $D = I(Z_1 + Z_2 > \nu)$, $T = \alpha(Z_1 + Z_2) + \epsilon_1$, and $\epsilon_1 \sim N(0, 0.25)$. We choose ν to make the disease prevalence equal to 0.3 and set $\pi(T, C) = \frac{e^{\delta_0+\delta_1 T+\delta_2 C}}{1+e^{\delta_0+\delta_1 T+\delta_2 C}}$, where $\delta_0 = 0.05$, $\delta_1 = 0.9$, $\delta_2 = 0.7$, and C is a binary variable obtained as $C = I(Z_1 + Z_2 + \epsilon_2 > 0)$, with $\epsilon_2 \sim N(0, 9)$. As for α, we maintain the values 0.1, 0.25, 0.5, 1 and 1.5, that give rise to the same five true AUC values 0.595, 0.714, 0.846, 0.943, 0.973. The number of replications in each simulation

Table 4. Study 3. Monte Carlo means (MC mean), Monte Carlo standard deviations (MC s.d.) and Monte Carlo means of bootstrap standard deviations (bootstrap s.d.) for the 1NN and 3NN AUC estimators (2) based on the Euclidean distance measure, for different values of α and different sample sizes. In the right side, empirical coverages of the confidence intervals for the AUC obtained through the normal approximation approach, with the standard deviation of $\widehat{AUC}^S$ estimated using the bootstrap method. Nominal coverages: 0.99, 0.95, 0.90

α	sample size		MC mean	MC s.d.	bootstrap s.d.	empirical coverages		
0.1	200	1NN	0.592	0.072	0.068	0.971	0.923	0.864
		3NN	0.591	0.065	0.066	0.976	0.937	0.893
	400	1NN	0.590	0.049	0.047	0.981	0.933	0.880
		3NN	0.590	0.045	0.046	0.986	0.947	0.905
0.25	200	1NN	0.708	0.064	0.060	0.974	0.922	0.867
		3NN	0.706	0.059	0.059	0.981	0.937	0.890
	400	1NN	0.707	0.044	0.042	0.983	0.935	0.880
		3NN	0.706	0.040	0.041	0.987	0.950	0.902
0.5	200	1NN	0.839	0.046	0.044	0.972	0.931	0.877
		3NN	0.836	0.043	0.043	0.980	0.941	0.897
	400	1NN	0.839	0.032	0.030	0.980	0.936	0.879
		3NN	0.838	0.029	0.030	0.983	0.946	0.896
1	200	1NN	0.940	0.024	0.023	0.956	0.914	0.867
		3NN	0.938	0.023	0.023	0.969	0.931	0.892
	400	1NN	0.940	0.016	0.016	0.976	0.934	0.883
		3NN	0.939	0.015	0.015	0.981	0.948	0.901
1.5	200	1NN	0.971	0.014	0.014	0.950	0.914	0.874
		3NN	0.970	0.013	0.014	0.961	0.930	0.895
	400	1NN	0.970	0.010	0.009	0.970	0.928	0.880
		3NN	0.970	0.009	0.009	0.975	0.942	0.898

experiment is 5000. We fix two sample sizes ($n = 200, 400$). For each Monte Carlo sample, we compute our 1NN and 3NN AUC estimates (2) based on the Euclidean distance measure with C as stratifying factor. Hence, we estimate the standard deviations of the corresponding estimators by 200 bootstrap samples.

Table 4 shows Monte Carlo means, Monte Carlo standard deviations and Monte Carlo means of bootstrap standard deviations for the 1NN and 3NN AUC estimators based on the Euclidean distance measure, for the chosen values of α and sample sizes. The table also reports the empirical coverages of the confidence intervals for the AUC obtained through the normal approximation approach, with the standard deviation of $\widehat{AUC}^S$ estimated using the bootstrap method. Confidence intervals have nominal coverages 0.99, 0.95, 0.90.

Results show a good behaviour of the estimators and of the boostrap procedure. Again, the 3NN AUC estimator achieves better results than the 1NN AUC estimator when the objective of inference is the construction of confidence intervals, in particular for low sample sizes. On the other hand, as in the previous study, we expect that the coverage accuracy may be generally improved on using the logit transformation.

Study 4. Monte Carlo experiments are used to compare, with respect to bias and standard deviation, the new method with the existing approaches MSI, IPW and SPE. We do not consider the FI method because of its similarities with the MSI method. Again, we focus our attention on 1NN and 3NN estimators based on the Euclidean distance measure, and we consider a vector $X = ({}_1X, {}_2X, {}_3X)^\top$ of three observed covariates. The number of replicates in each simulation experiment is 5000.

The MSI method is based on a parametric model for $\rho(y)$, the IPW method is based on a parametric model for $\pi(y)$, and the SPE method is based on both models. Clearly, a

Table 5. Study 4 (i). Monte Carlo means and standard deviations of the KNN AUC estimators and competitors, for different values of α. X has dimension 3. Models for $\rho(y)$ and $\pi(y)$, chosen to obtain MSI, IPW and SPE estimators, are both correctly specified

	$\alpha = 0.1$		$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 1.5$	
	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.
sample size = 200										
Full	0.627	0.043	0.774	0.035	0.897	0.022	0.967	0.011	0.985	0.006
Naïve	0.569	0.058	0.708	0.051	0.843	0.038	0.937	0.023	0.965	0.015
1NN	0.624	0.047	0.769	0.039	0.891	0.027	0.963	0.014	0.982	0.009
3NN	0.622	0.044	0.766	0.037	0.887	0.025	0.960	0.014	0.980	0.009
MSI	0.626	0.047	0.774	0.038	0.897	0.025	0.967	0.012	0.985	0.007
IPW	0.622	0.067	0.770	0.055	0.893	0.034	0.964	0.016	0.982	0.010
SPE	0.627	0.048	0.774	0.039	0.897	0.026	0.967	0.012	0.984	0.007
sample size = 500										
Full	0.627	0.026	0.774	0.022	0.897	0.014	0.967	0.007	0.984	0.004
Naïve	0.571	0.036	0.708	0.032	0.843	0.024	0.937	0.014	0.965	0.010
1NN	0.625	0.030	0.770	0.024	0.893	0.017	0.964	0.009	0.983	0.005
3NN	0.624	0.028	0.769	0.023	0.891	0.016	0.963	0.008	0.982	0.005
MSI	0.627	0.029	0.774	0.024	0.897	0.016	0.967	0.007	0.984	0.004
IPW	0.626	0.041	0.772	0.034	0.895	0.022	0.966	0.010	0.983	0.006
SPE	0.627	0.030	0.774	0.024	0.897	0.016	0.967	0.008	0.984	0.004

wrong specification of such models may affect the estimation. Hence, in this study we consider the following scenarios: (i) the models for $\rho(y)$ and $\pi(y)$, chosen to obtain MSI, IPW and SPE estimators, are both correctly specified; (ii) the models for $\rho(y)$ and $\pi(y)$, chosen to obtain MSI, IPW and SPE estimators, are both misspecified. Scenario (i) allows to evaluate the behaviour of the KNN estimators in settings 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.

Moreover, we consider also another scenario: (iii) in the estimation process, a covariate which is not involved neither in the disease nor in the verification processes is introduced and a relevant covariate is omitted. Clearly, this scenario allows us to evaluate possible effects of a particular kind of misspecification on all estimators (including KNN estimators).

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

The simulation setting is a generalization of that of Study 1. Starting from four independent random variables, $Z_1 \sim N(0, 0.5)$ to $Z_4 \sim N(0, 0.5)$, the disease indicator D is specified as $D = I[Z_1 + Z_2 + Z_3 + Z_4 > \nu]$. The threshold ν determines a disease prevalence of about 0.31. The diagnostic test result T and the auxiliary covariates are generated as follows:

- $T = \alpha \sum_{i=1}^{4} Z_i + \epsilon_1$,
- ${}_1X = 0.1 \sum_{i=1}^{4} Z_i + \epsilon_2$,
- ${}_2X = 0.5Z_1 + 2Z_2 + 1.5Z_3 + 3Z_4 + \epsilon_3$,
- ${}_3X = 2Z_1 - 0.5Z_2 + Z_3 + 0.5Z_4 + \epsilon_4$,

where ϵ_i, $i = 1, \dots 4$, are independent $N(0, 0.25)$ random variables, independent also from Z_i, $i = 1, \dots, 4$. In accordance with the MAR assumption, the verification probability π is set to be

$$\pi(T, X) = \frac{e^{\delta_0 + \delta_1 T + \delta_2^\top X}}{1 + e^{\delta_0 + \delta_1 T + \delta_2^\top X}},$$

with $\delta_0 = 0.05$, $\delta_1 = 0.9$, $\delta_2 = (0.7, 0.4, 0.2)^\top$. This choice corresponds to a verification rate of about 0.55. Finally, as for α, we again choose the values 0.1, 0.25, 0.5, 1 and 1.5, that give rise to values of the true AUC ranging between 0.62 and 0.85, approximately. We fix two sample sizes: a moderate one, i.e., $n = 200$, and a relatively high one, i.e., $n = 500$. For the IPW, MSI and SPE estimators, conditional disease probabilities and conditional verification probabilities are estimated using correctly specified models. More precisely, we use a generalized linear model for D given T and X with probit link (see [2]). The conditional verification probabilities are estimated from a logistic regression model with V as the response and T and X as predictors.

Table 5 shows Monte Carlo means and standard deviations of the AUC estimators. Results concern the estimators IPW, MSI, SPE and the new proposals 1NN and 3NN based on the Euclidean distance. Rows denoted by "Naïve" indicate the results for the Wilcoxon statistic (the nonparametric AUC estimator) computed by using the verified cases only. Again, the Full estimator is used as benchmark.

From the simulation results it is clear that all (partially) parametric methods behave well if models for $\rho(y)$ and $\pi(y)$ are both correctly specified, with the IPW method showing

Table 6. Study 4 (ii). Monte Carlo means and standard deviations of the KNN AUC estimators and competitors, for different paires of (α, β). X has dimension 3. The models for $\rho(y)$ and $\pi(y)$, chosen to obtain MSI, IPW and SPE estimators, are both misspecified. Sample size $= 500$. The SPE estimator produces estimates outside of the $[0, 1]$ interval with rate equal to 6.5%, 22%, 65% and 25% when the pair (α, β) is $(0.15, 3)$, $(0.5, 2)$, $(1.5, 1)$ and $(5, 0.6)$, respectively. The row denoted by "SPE" gives the results for the SPE estimator when the estimates outside of the $[0, 1]$ interval are truncated to be either 0 or 1*

	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.
	$\alpha = 0.15$, $\beta = 3$		$\alpha = 0.5$, $\beta = 2$		$\alpha = 1.5$, $\beta = 1$		$\alpha = 5$, $\beta = 0.6$	
Full	0.608	0.025	0.713	0.023	0.848	0.017	0.929	0.011
Naïve	0.552	0.064	0.628	0.063	0.766	0.051	0.866	0.036
1NN	0.624	0.052	0.713	0.058	0.831	0.051	0.920	0.030
3NN	0.625	0.041	0.713	0.050	0.822	0.048	0.911	0.030
MSI	0.516	0.054	0.577	0.054	0.703	0.053	0.864	0.033
IPW	0.655	0.114	0.774	0.097	0.906	0.062	0.955	0.029
SPE	0.658	1.849	0.922	2.947	1.186	5.546	0.972	0.065
SPE*	0.681	0.155	0.821	0.154	0.944	0.158	0.960	0.041

slightly poorer performances in some circumstances. However, in terms of bias and standard deviation, the new proposals compare very well with existing estimators, and the estimators 1NN and 3NN seem to achieve similar performances.

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

Starting from four independent random variables Z_1 to Z_4, such that $Z_i/\sqrt{0.5} \sim EXP(1)$, the disease indicator D is specified as $D = I[Z_1 Z_2 + Z_4 > \nu]$. The threshold ν determines a disease prevalence of about 0.38. The diagnostic test result T and the auxiliary covariates are generated as follows:

- $T = \alpha(Z_1 Z_2 + \beta Z_3 + Z_4) + \epsilon_1$,
- ${}_1X = 0.5\left(\sum_{i=1}^{4} Z_i\right)^2 + \epsilon_2$,
- ${}_2X = 0.5Z_1^2 + 2Z_2^2 + 1.5Z_3^2 + 3Z_4^2 + \epsilon_3$,
- ${}_3X = 2Z_1^2 - 0.5Z_2^2 + Z_3^2 + 0.5Z_4^2 + \epsilon_4$,

where $\epsilon_i/\sqrt{0.25}$, $i = 1, \ldots 4$, are independent $EXP(1)$ random variables, independent also from Z_i, $i = 1, \ldots, 4$. We consider four values for the pair (α, β), i.e., (0.15,3), (0.5,2), (1.5,1) and (5,0.6) giving rise to four different true AUC values. Finally, the verification probability π is set to be

$$\pi(T, X) = 0.05 + 0.1I[T > 2,\ {}_2X > 2] + 0.85I[{}_1X {}_3X > 2].$$

This choice corresponds to a verification rate of about 0.2.

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. 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 subjects. This has led us to the choice of $n = 500$.

For the (partially) parametric estimators IPW, MSI, SPE, to estimate the conditional disease probabilities, we use a generalized linear model for D given T and X with logit link; this model is clearly 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.

Results are given in Table 6. They evidently show a very poor behaviour of the estimators MSI, IPW and SPE. In particular, the performance of the SPE estimator is surprisingly negative, providing a high number of estimates outside of the $[0, 1]$ interval. Our 1NN and 3NN estimators, instead, show a good behaviour. Moreover, it is worth noting that, in this setting, the function $\pi(y)$ used to mimic the verification process is not smooth. Hence, the KNN AUC estimators seem to show also some degree of robustness against violation of smoothness assumptions.

(iii) Substituting a relevant variable with an independent one.

The simulation setting is the same as that of Study 4(i). There is only one difference: at each simulation run, in addition to the random variables Z_1, Z_2, Z_3 and Z_4, we generate also another variable $Z_5 \sim N(0, 0.5)$, independent of all others. Hence, the AUC estimates are obtained substituting in turn Z_5 to ${}_1X$, ${}_2X$ and ${}_3X$ in the estimation process. Sample size is $n = 200$.

Simulation results are given in Table 7. Based on these results, the effect of such a misspecification seems to be generally weak, and slightly more significant for the KNN estimators. As expected, however, for all estimators the worst results occur when the missing covariate in the estimation process is ${}_2X$, i.e. the variable more strongly associated with D and V, and the correlation between T and ${}_2X$ is low (approximately 0.24, when $\alpha = 0.1$, and 0.51, when $\alpha = 0.25$).

4. AN ILLUSTRATION

To illustrate the application of the proposed method, we used a data set within the Uniform Data Set (UDS) of

Table 7. Study 4 (iii). Monte Carlo means and standard deviations of the KNN AUC estimators and competitors, for different values of α. X has dimension 3. Sample size $= 200$. Estimators are computed using Z_5 in place of one covariate ($_1X$, $_2X$ or $_3X$) belonging to the data generator model

	$\alpha = 0.1$		$\alpha = 0.25$		$\alpha = 0.5$		$\alpha = 1.0$		$\alpha = 1.5$	
	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.	MC mean	MC s.d.
Full	0.627	0.043	0.774	0.035	0.897	0.022	0.967	0.011	0.985	0.006
Z_5 instead of $_1X$										
Naïve	0.571	0.057	0.708	0.052	0.844	0.039	0.937	0.023	0.965	0.015
1NN	0.624	0.047	0.768	0.039	0.890	0.027	0.962	0.014	0.981	0.009
3NN	0.623	0.044	0.765	0.037	0.887	0.025	0.959	0.015	0.980	0.009
MSI	0.627	0.047	0.774	0.039	0.898	0.025	0.967	0.012	0.985	0.007
IPW	0.625	0.065	0.769	0.053	0.894	0.034	0.965	0.016	0.982	0.009
SPE	0.628	0.047	0.774	0.040	0.897	0.026	0.967	0.012	0.985	0.007
Z_5 instead of $_2X$										
Naïve	0.571	0.058	0.708	0.052	0.843	0.038	0.936	0.022	0.965	0.015
1NN	0.584	0.058	0.717	0.056	0.852	0.045	0.953	0.021	0.979	0.012
3NN	0.581	0.050	0.708	0.049	0.839	0.041	0.945	0.022	0.976	0.012
MSI	0.595	0.055	0.744	0.049	0.883	0.031	0.964	0.013	0.984	0.008
IPW	0.591	0.068	0.740	0.063	0.878	0.043	0.962	0.018	0.981	0.010
SPE	0.593	0.062	0.743	0.057	0.882	0.037	0.964	0.014	0.984	0.008
Z_5 instead of $_3X$										
Naïve	0.570	0.056	0.708	0.051	0.843	0.039	0.937	0.022	0.966	0.015
1NN	0.617	0.049	0.760	0.044	0.883	0.033	0.961	0.017	0.982	0.009
3NN	0.615	0.045	0.756	0.041	0.878	0.031	0.957	0.018	0.980	0.011
MSI	0.621	0.048	0.769	0.041	0.895	0.027	0.967	0.012	0.985	0.007
IPW	0.617	0.068	0.764	0.056	0.891	0.035	0.964	0.017	0.982	0.010
SPE	0.621	0.051	0.769	0.043	0.894	0.029	0.966	0.013	0.984	0.007

National Alzheimer's Coordinating Center (NACC), which came from 32 Alzheimer's Disease Centers throughout North America since 2006. The patients were referred or self-referred for evaluation of possible dementia, or recruited specifically to participate in clinical research. Most patients underwent clinical evaluation and neuropsychological tests for cognitive impairment at enrollment. During the follow-up period, the patients received periodical re-evaluation and cognitive tests. Among these cognitive tests, the mini-mental state examination (MMSE) is a brief 30-point questionnaire test that is widely used to screen for cognitive impairment. In general, scores of 27 or above (out of 30) are considered normal. Although the MMSE score is measured on a discrete scale, in medical studies is quite commonly treated as a continuous measurement (see, for example, [8]). In the progression of dementia, the amnestic mild cognitive impairment (aMCI) is an important transitional stage. Patients with aMCI could still revert to normal, but dementia is generally believed to be irreversible.

[7] previously used this data set to investigate the one-year progression from aMCI to dementia, and find out how well the baseline MMSE score classifies the patients who progressed to dementia and those who did not in one year. The authors included in the study patients who aged over 65 and were diagnosed to be aMCI at their first visit. If a patient made a visit about one year (within the 6–18 months window) after the baseline, his/her cognitive status is observed, with $D = 1$ indicating progression to dementia and 0 otherwise. The disease status was missing if the patient only made the baseline visit, or the follow-up visits were all outside the 6–18 months window. The covariates used by [7] in the ROC analysis included age, gender, race, marital status, living situation, stroke, and history of cardiovascular diseases. Other disease history variables, and clinical dementia rating (CDR) sum of boxes were considered as the predictor for the missingness mechanism and the disease model. Subjects with missing covariates were excluded. Relevant conclusions emerged from the study were as follows. (a) The progression of dementia is complicated and not fully understood, and in this study, the missingness could be due to various reasons; hence, the SPE method is recommended in this example, which protects the model misspecification under the MAR assumption. (b) The SPE AUC estimator showed that MMSE has some classification accuracy only for patients with no stroke and with more than 17 years of education.

In light of statement (a), we decided that it was important to check the accuracy of the MMSE score in the subgroup of patients specified in (b) by using a fully nonparametric approach. To this end, we used the KNN method to estimate

the AUC of the MMSE score, for patients with more than 17 years of education and without stroke. The sample for the analysis consisted of 975 subjects, for 595 of which the cognitive status was observed at the second visit. Within the verified sample, the progression to dementia was observed 148 times (approximately 25%). To obtain KNN AUC estimates, we included in the disease model the MMSE score (T), the disease history information (CDR sum of boxes), the age of onset of cognitive decline and an artificial real valued random component (generated from a standard normal variate) to break distance ties. Due to the MMSE measuring scale, 1NN and 3NN AUC estimators were computed by using formula (3) (after changing the sign of T), which accommodates for the presence of score ties. By using the Euclidean distance, we obtained the estimates 0.708 (1NN) and 0.705 (3NN), with estimated standard deviations equal to 0.025 and 0.023, respectively (95% CI: 0.658, 0.758 and 0.659, 0.752). Using the Mahalanobis distance, instead, we obtained the estimates 0.697 (1NN) and 0.698 (3NN), with estimated standard deviations equal to 0.026 (95% CI: 0.645, 0.748 and 0.647, 0.750). Therefore, our results seem to confirm that the MMSE score is not a satisfactory marker for predicting progression to dementia, and suggest caution in its use.

A final remark concerns the computational burden. In this example, having a sample size of about 1000, a dimension of Y equal to 4 and a missingness rate of about 39%, with a fast matrix programming language like GAUSS, the estimates are computed in few seconds, and approximately 40 seconds are required to get confidence intervals (with 200 bootstrap replications).

5. DISCUSSION

In this paper we have developed an approach to verification bias-corrected inference on the AUC of a continuous-scale diagnostic test, that does not rely on parametric assumptions about the disease and/or selection models. The approach foresees the presence of continuous covariates and it is naturally extended to stratified samples in which strata are defined according to the presence of categorical variables. The proposed approach works under MAR assumption.

The new AUC estimators are fully nonparametric. They represent an alternative to the classic (partially) parametric estimators, and their use can reduce the effects of possible misspecifications to the inference, as shown by the results of our simulation Study 4(ii). Of course, if one uses the new estimators in situations where the disease and the verification models can be correctly specified, one expects a loss of efficiency compared to the use of parametric alternatives. Such loss may be small, as show, for example, in our Study 4(i), but it may certainly be significant in certain circumstances, especially when the sample size is small.

The new method is based on the K-nearest-neighbor imputation, which requires the choice of a value for K as well as a distance measure. In our simulation Study 1, performance of the KNN estimators are quite comparable for different choices of the distance measure. In practical situations, however, the selection of a suitable distance is generally dictated by features of the data, possible subjective evaluations and by computational concerns, so that a general indication on an adequate choice is difficult to express. As for the choice of the size of the neighbor, our simulation results suggest that a value for K around 3 could to be adequate as long as the dimension of Y is not large. A similar conclusion also arises in [1], where K-nearest-neighbor imputation is applied to estimate a ROC curve and in [9], where K-nearest-neighbor imputation is applied to estimate a mean functional. However, if the dimension of Y increases, it could be convenient to consider higher values for K. The selected value must be compatible with (i.e. not too big with respect to) the number of verified units present in the sample. Generally speaking, a possible strategy to choose a suitable value for K in practice could be cross-validation, as illustrated in [1].

Standard deviation of our KNN estimators can be estimated by a simple bootstrap procedure. Our simulation results show effectiveness of this procedure. Nevertheless, due to its fully nonparametric nature, the proposed approach requires sufficient information from the data to provide accurate inference on the AUC. In particular, as indicated by our simulation results, when the objective is to build confidence intervals, we expect that the sample size needed to achieve sufficient accuracy may depend on the true AUC value and on the rate of verified units (healthy as well as diseased) in the sample. High values of AUC and small verification rates will likely require a high sample size. From a practical point of view, however, to improve the accuracy of the confidence intervals obtained using the approach proposed in the paper, we strongly recommend the use of the logit transformation.

We are aware that, when auxiliary data X come in the form of a high-dimensional feature vector, appeal of our estimators diminishes, at least from a technical point of view. Firstly, the nearest neighbors search is computationally demanding. More generally, the dimensionality curse phenomenon states that in high dimensional spaces distances between nearest and farthest observations from a given subject become almost equal. This is often cited as "distance functions losing their usefulness in high dimensionality". In the data mining literature, various solutions are proposed both to cut the computational costs of the nearest neighbors search and to provide measures of similarities between data points able to better grasp divergence between the maximum and minimum distances (see, for example, [11] as a general reference). However, from a classical statistical perspective, relevance of such drawbacks is softened by the consideration that, in practice, analysts modeling complex data rarely use all available information. Therefore, a sound solution to problems caused by high dimensionality relies on a screening and selection of the auxiliary variables, see [3].

We conclude with a last remark. Often, the auxiliary covariate vector X is directly associated with the marker T under study. In this case, one can be interested in studying the covariate-specific ROC curves and the related covariate-specific AUCs (see, for example, [7]). We believe that the KNN strategy could be usefully employed in this context and plan future research on such theme.

ACKNOWLEDGEMENTS

The authors wish to thank National Alzheimer's Coordinating Center (NACC) for providing the data for analysis and the Referees for their helpful comments on an earlier version of the paper.

Received 15 September 2015

REFERENCES

[1] Adimari, G. and Chiogna, M. (2015). Nearest-neighbor estimation for ROC analysis under verification bias. *The International Journal of Biostatistics*, **11**, 109–124. MR3341515

[2] Alonzo, T. A. and Pepe, M. S. (2005). Assessing accuracy of a continuous screening test in the presence of verification bias. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, **54**(1), 173–190. MR2134605

[3] Budczies, J., Kosztyla, D., von Törne, C., Stenzinger, A., Darb-Esfahani, S., Dietel, M., and Denkert, C. (2014). Cancerclass: An R package for development and validation of diagnostic tests from high-dimensional molecular data. *Journal of Statistical Software*, **59**.

[4] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabilistic Theory of Pattern Recognition*. Springer. MR1383093

[5] He, H., Lyness, J. M., and McDermott, M. P. (2009). Direct estimation of the area under the receiver operating characteristic curve in the presence of verification bias. *Statistics in Medicine*, **28**, 361–376. MR2655685

[6] Liu, D. and Zhou, X. H. (2010). A model for adjusting for non-ignorable verification bias in estimation of the ROC curve and its area with likelihood-based approach. *Biometrics*, **66**, 1119–1128. MR2758499

[7] Liu, D. and Zhou, X.-H. (2013). Covariate adjustment in estimating the area under ROC curve with partially missing gold standard. *Biometrics*, **69**, 91–100. MR3058055

[8] Newsom, J., Jones, R. N., and Hofer, S. M. (Editor) (2012). *Longitudinal Data Analysis: A Practical Guide for Researchers in Aging, Health, and Social Sciences*. Routledge, Taylor & Francis Group.

[9] Ning, J. and Cheng, P. E. (2012). A comparison study of non-parametric imputation methods. *Statistics and Computing*, **22**, 273–285. MR2865070

[10] Rotnitzky, A., Faraggi, D., and Schisterman, E. (2006). Doubly robust estimation of the area under the receiver-operating characteristic curve in the presence of verification bias. *Journal of the American Statistical Association*, **101**, 1276–1288. MR2328313

[11] Shakhnarovich, G., Darrell, T., and Indyk, P. (2006). *Nearest-Neighbor Methods in Learning and Vision: Theory and Practice*. The MIT Press.

Gianfranco Adimari
Department of Statistical Sciences
University of Padova
Via C. Battisti, 241–243
35121 Padova
Italy
E-mail address: gianfranco.adimari@unipd.it

Monica Chiogna
Department of Statistical Sciences
University of Padova
Via C. Battisti, 241–243
35121 Padova
Italy
E-mail address: monica.chiogna@unipd.it