REVSTAT – Statistical Journal
Volume 18, Number 5, October 2020, 697–720

NONPARAMETRIC ESTIMATION OF ROC SURFACES UNDER VERIFICATION BIAS

Authors: Khanh To Duc
– Department of Statistical Sciences, University of Padova, Via C. Battisti, 241-243, 35121 Padova, Italy
toduc@stat.unipd.it

Monica Chiogna
– Department of Statistical Sciences, University of Padova, Via C. Battisti, 241-243, 35121 Padova, Italy
monica@stat.unipd.it

Gianfranco Adimari
– Department of Statistical Sciences, University of Padova, Via C. Battisti, 241-243, 35121 Padova, Italy
adimari@stat.unipd.it

Received: February 2017 Revised: April 2018 Accepted: August 2018

Abstract:

- Verification bias is a well known problem that can affect the statistical evaluation of the predictive ability of a diagnostic test when the true disease status is unknown for some of the patients under study. In this paper, we deal with the assessment of continuous diagnostic tests when an (ordinal) three-class disease status is considered and propose a fully nonparametric verification bias-corrected estimator of the ROC surface based on nearest-neighbor imputation. Consistency and asymptotic normality of the proposed estimator are proved under the missing at random assumption, and its finite sample behavior is investigated by means of Monte Carlo experiments. Variance estimation is also discussed and an illustrative example is presented.

Key-Words:

- *diagnostic tests; missing at random; true class fractions; nearest neighbor imputation.*

AMS Subject Classification:

- 62C99, 62P10.

1. INTRODUCTION

The assessment of diagnostic tests is an important issue in modern medicine. In a two-class problem, i.e. when the disease status has two categories (e.g., "healthy" and "diseased"), for a diagnostic test T that yields a continuous measure, the receiver operating characteristic (ROC) curve is a popular tool for displaying the ability of the test to distinguish between the classes. Assuming, without loss of generality, that higher test values indicate a higher likelihood of disease, the ROC curve is defined as the set of points $\{(1-\mathrm{TNR}(c), \mathrm{TPR}(c)), c \in (-\infty, \infty)\}$ in the unit square, where c is a cut point value, $\mathrm{TPR}(c) = \Pr(T \geq c \mid \text{subject is diseased})$ is the true positive rate at c and $\mathrm{TNR}(c) = \Pr(T < c \mid \text{subject is non–diseased})$ is the true negative rate at c. The shape of the ROC curve allows to evaluate the ability of the test. For example, a ROC curve equal to a straight line joining points $(0,0)$ and $(1,1)$ represents a diagnostic test which is the random guess. A commonly used summary measure of the overall performance of the test is the area under ROC curve (AUC). Under correct ordering, values of AUC range from 0.5, suggesting that the test is no better than chance alone, to 1.0, which indicates a perfect test. See, for example, [13] and [17] as general references.

In some medical studies, the disease status often involves three classes; see, for example, [5], [6] and [11]. In such situations, quantities used to evaluate the accuracy of tests are the true class fractions (TCF's). These quantities are defined as generalizations of TPR and TNR. For a given pair of cut points (c_1, c_2) such that $c_1 < c_2$, the true class fractions TCF's of the continuous test T at (c_1, c_2) are

$$\begin{aligned}
\mathrm{TCF}_1(c_1) &= \Pr(T < c_1 | \text{class } 1) = 1 - \Pr(T \geq c_1 | \text{class } 1), \\
\mathrm{TCF}_2(c_1, c_2) &= \Pr(c_1 \leq T < c_2 | \text{class } 2) \\
&= \Pr(T \geq c_1 | \text{class } 2) - \Pr(T \geq c_2 | \text{class } 2), \\
\mathrm{TCF}_3(c_2) &= \Pr(T \geq c_2 | \text{class } 3) = \Pr(T \geq c_2 | \text{class } 3).
\end{aligned}$$

The plot of (TCF_1, TCF_2, TCF_3) at various values of the pair (c_1, c_2) produces the ROC surface, a generalization of the ROC curve to the unit cube (see [11],[10],[15]). The ROC surface is the region defined by the triangle with vertices $(0,0,1), (0,1,0)$, and $(1,0,0)$ if the three TCF's are identical for every pair (c_1, c_2). In this case, we say that the diagnostic test is, again, the random guess. The ROC surface of an effective test lies in the unit cube above such region. A summary measure of the overall diagnostic accuracy of the test under consideration is the volume under the ROC surface (VUS), which can be seen as a generalization of the AUC. For correctly ordered categories, values of VUS vary from 1/6 to 1, ranging from bad to perfect diagnostic tests.

The application of a diagnostic test in the clinical practice requires a preliminary rigorous statistical assessment of its performance. Clearly, the true ROC curve (or surface) of the test under assessment and its AUC (or VUS) are unknown, so that the statistical evaluation relies on suitable inferential procedures, typically based on measurements collected on a sample of patients. The assessment requires to ascertain the true disease status of the patients in the sample, a verification that it is generally done by employing the most accurate available test, the so-called gold standard (GS) test. Some times, however, the GS test is too expensive, or too invasive, or both to be used on large samples, so that only a subset

of patients undergoes disease verification. It happens that statistical evaluations based only on data from subjects with verified disease status are typically biased, an effect known as verification bias.

Correcting for verification bias is a well known issue of medical statistics. Various methods have been developed to deal with the problem, most of which refer to the two-class case and assume that the true disease status, if missing, is missing at random (MAR, see [9]). We recall, among others, papers [1], [2], [3], [7], [14] and [17]. In particular, for continuous tests, [3] proposes four types of partially parametric estimators of TPR and TNR under the MAR assumption, i.e., full imputation (FI), mean score imputation (MSI), inverse probability weighting (IPW) and semiparametric efficient (SPE, also known as doubly robust DR) estimators. [1] and [2], instead, propose a fully nonparametric approach for ROC curve and AUC estimation, respectively.

The issue of correcting for verification bias in ROC surface analysis is very scarcely considered in the literature. To the best of our knowledge, only [5] and [16] discuss the issue. [5] proposes a maximum likelihood approach for estimation of the ROC surface and corresponding VUS for ordinal diagnostic tests, whereas [16] extends methods in [3] to the estimation of ROC surfaces of continuous diagnostic tests. It is worth noting that FI, MSI, IPW and SPE estimators in [16] are partially parametric estimators and their use requires the specification of parametric regression models for the probability of a subject being correctly classified with respect to the disease state, or the probability of a subject being verified (i.e., tested by GS), or both. As a consequence, a wrong specification of such parametric models negatively affects the behavior of the estimators, that no longer are consistent.

To avoid problems due to model misspecification, in this paper we propose a fully nonparametric approach to estimate TCF_1, TCF_2 and TCF_3 in the presence of verification bias, for continuous diagnostic tests. The proposed approach is based on a nearest-neighbor (NN) imputation of the missing data and extends an idea developed in [1]. Consistency and asymptotic normality of the estimators derived from the proposed method are studied. In addition, estimation of their variance is also discussed. Usefulness of our proposal and advantages in comparison with partially parametric estimators is assessed with the aid of some simulation experiments. An illustrative example is also given.

The rest of paper is organized as follows. In Section 2, we review partially parametric methods for correcting for verification bias in case of continuous tests. The proposed nonparametric method for (pointwise) estimating ROC surfaces and the related asymptotic results are presented in Section 3. In Section 4, we discuss variance-covariance estimation and in Section 5 we give some simulation results. An application is illustrated in Section 6. Finally, conclusions are drawn in Section 7. Some technical details and other simulation results are available in a Supplementary Material, downloadable at http://paduaresearch.cab.unipd.it/11221/ .

2. PARTIALLY PARAMETRIC ESTIMATORS OF ROC SURFACES

Consider a study with n subjects, for whom the result of a continuous diagnostic test T is available. For each subject, $\mathcal{D}$ denotes the true disease status, that can possibly be unknown. Hereafter, we will describe the true disease status as a trinomial random vector $\mathcal{D} = (D_1, D_2, D_3)$. D_k is a binary variable that takes 1 if the subject belongs to class k, $k = 1, 2, 3$ and 0 otherwise. Here, class 1, class 2 and class 3 can be referred, for example, as "non-diseased", "intermediate" and "diseased", and are assumed to be ordered. Further, let V be a binary verification status for a subject, such that $V = 1$ if he/she is undergoes the GS test, and $V = 0$ otherwise. In practice, some information, other than the results from the test T, can be obtained for each patient. Let A be the covariate vector for the patients, that may be associated both with $\mathcal{D}$ and V. We are interested in estimating the ROC surface of T, and hence the true class fractions $\mathrm{TCF}_1(c_1) = \Pr(T_i < c_1 | D_{1i} = 1)$, $\mathrm{TCF}_2(c_1, c_2) = \Pr(c_1 \leq T_i < c_2 | D_{2i} = 1)$ and $\mathrm{TCF}_3(c_2) = \Pr(T_i \geq c_2 | D_{3i} = 1)$, for fixed constants c_1, c_2, with $c_1 < c_2$.

When all patients have their disease status verified by a GS, i.e., $V_i = 1$ for all $i = 1, ..., n$, for any pair of cut points (c_1, c_2), the true class fractions $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ can be easily estimated by

$$\widehat{\mathrm{TCF}}_1(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) D_{1i}}{\sum_{i=1}^{n} D_{1i}},$$

$$\widehat{\mathrm{TCF}}_2(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) D_{2i}}{\sum_{i=1}^{n} D_{2i}},$$

$$\widehat{\mathrm{TCF}}_3(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) D_{3i}}{\sum_{i=1}^{n} D_{3i}},$$

where $\mathrm{I}(\cdot)$ is the indicator function. It is straightforward to show that the above estimators are consistent. However, they cannot be employed in case of incomplete data, i.e. when $V_i = 0$ for some $i = 1, ..., n$.

When only some subjects are selected to undergo the GS test, we need to make an assumption about the selection mechanism. We assume that the verification status V and the disease status $\mathcal{D}$ are mutually independent given the test result T and covariate A. This means that $\Pr(V|T, A) = \Pr(V|\mathcal{D}, T, A)$ or equivalently $\Pr(\mathcal{D}|T, A) = \Pr(\mathcal{D}|V, T, A)$. Such assumption is a special case of the missing at random (MAR) assumption (see [9]).

Under MAR assumption, verification bias-corrected estimation of the true class fractions is discussed in [16], where (partially) parametric estimators, based on four different approaches, are given. In particular, full imputation (FI) estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$

and $\mathrm{TCF}_3(c_2)$ are defined as

$$\widehat{\mathrm{TCF}}_{1,\mathrm{FI}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1)\hat{\rho}_{1i}}{\sum_{i=1}^{n} \hat{\rho}_{1i}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{FI}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2)\hat{\rho}_{2i}}{\sum_{i=1}^{n} \hat{\rho}_{2i}}, \tag{2.1}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{FI}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2)\hat{\rho}_{3i}}{\sum_{i=1}^{n} \hat{\rho}_{3i}}.$$

This method requires a parametric model (e.g. multinomial logistic regression model) to obtain the estimates $\hat{\rho}_{ki}$ of $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$, using only data from verified subjects. Differently, the mean score imputation (MSI) approach only uses the estimates $\hat{\rho}_{ki}$ for the missing values of disease status D_{ki}. Hence, MSI estimators are

$$\widehat{\mathrm{TCF}}_{1,\mathrm{MSI}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1)\left[V_i D_{1i} + (1 - V_i)\hat{\rho}_{1i}\right]}{\sum_{i=1}^{n} \left[V_i D_{1i} + (1 - V_i)\hat{\rho}_{1i}\right]},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{MSI}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2)\left[V_i D_{2i} + (1 - V_i)\hat{\rho}_{2i}\right]}{\sum_{i=1}^{n} \left[V_i D_{2i} + (1 - V_i)\hat{\rho}_{2i}\right]}, \tag{2.2}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{MSI}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2)\left[V_i D_{3i} + (1 - V_i)\hat{\rho}_{3i}\right]}{\sum_{i=1}^{n} \left[V_i D_{3i} + (1 - V_i)\hat{\rho}_{3i}\right]}.$$

The inverse probability weighting (IPW) approach weights each verified subject by the inverse of the probability that the subject is selected for verification. Thus, $\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ are estimated by

$$\widehat{\mathrm{TCF}}_{1,\mathrm{IPW}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) V_i \hat{\pi}_i^{-1} D_{1i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{1i}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{IPW}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) V_i \hat{\pi}_i^{-1} D_{2i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{2i}}, \tag{2.3}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{IPW}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) V_i \hat{\pi}_i^{-1} D_{3i}}{\sum_{i=1}^{n} V_i \hat{\pi}_i^{-1} D_{3i}},$$

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

$$\widehat{\mathrm{TCF}}_{1,\mathrm{SPE}}(c_1) = 1 - \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_1) \left\{ \frac{V_i D_{1i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{1i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_{1i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{1i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}},$$

$$\widehat{\mathrm{TCF}}_{2,\mathrm{SPE}}(c_1, c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) \left\{ \frac{V_i D_{2i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{2i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_{2i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{2i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}, \tag{2.4}$$

$$\widehat{\mathrm{TCF}}_{3,\mathrm{SPE}}(c_2) = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) \left\{ \frac{V_i D_{3i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{3i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}{\sum_{i=1}^{n} \left\{ \frac{V_i D_{3i}}{\hat{\pi}_i} - \frac{\hat{\rho}_{3i}(V_i - \hat{\pi}_i)}{\hat{\pi}_i} \right\}}.$$

Estimators (2.1)-(2.4) represent an extension to the three-classes problem of the estimators proposed in [3]. SPE estimators are also known to be doubly robust estimators, in the sense that they are consistent if either the ρ_{ki}'s or the π_i's are estimated consistently. However, SPE estimates could fall outside the interval $(0, 1)$. This happens because the quantities $V_i D_{ki} \hat{\pi}_i^{-1} - \hat{\rho}_{ki}(V_i - \hat{\pi}_i)\hat{\pi}_i^{-1}$ can be negative.

3. NONPARAMETRIC ESTIMATORS

3.1. The proposed method

All the verification bias-corrected estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ revised in the previous section belong to the class of (partially) parametric estimators, i.e., they need regression models to estimate $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$ and/or $\pi_i = \Pr(V_i = 1|T_i, A_i)$. In what follows, we propose a fully nonparametric approach to the estimation of $\mathrm{TCF}_1(c_1), \mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$. Our approach is based on the K-nearest neighbor (KNN) imputation method. Hereafter, we shall assume that A is a continuous random variable.

Recall that the true disease status is a trinomial random vector $\mathcal{D} = (D_1, D_2, D_3)$ such that D_k is a n Bernoulli trials with success probability $\theta_k = \Pr(D_k = 1)$. Note that $\theta_1 + \theta_2 + \theta_3 = 1$. Since parameters θ_k are the means of the random variables D_k, we can use the KNN estimation procedure discussed in [12] to obtain nonparametric estimates $\hat{\theta}_{k,\mathrm{KNN}}$. More precisely, we define

$$\hat{\theta}_{k,\mathrm{KNN}} = \frac{1}{n} \sum_{i=1}^{n} \left[V_i D_{ki} + (1 - V_i)\hat{\rho}_{ki,K} \right], \qquad K \in \{1, 2, 3, ...\},$$

where $\hat{\rho}_{ki,K} = \frac{1}{K} \sum_{l=1}^{K} D_{ki(l)}$, and $\left\{ (T_{i(l)}, A_{i(l)}, D_{ki(l)}) : V_{i(l)} = 1, l = 1, ..., K \right\}$ is a set of K observed data triplets and $(T_{i(l)}, A_{i(l)})$ denotes the l-th nearest neighbor to (T_i, A_i) among all (T, A)'s corresponding to verified patients, i.e., patients with $V = 1$.

Let $\beta_{jk} = \Pr(T \geq c_j, D_k = 1)$, with $j \in \{1, 2\}$, $k \in \{1, 2, 3\}$ and $k \geq j$. Then, we can define the KNN estimates of β_{jk} as

$$\hat{\beta}_{jk,\mathrm{KNN}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) \left[V_i D_{ki} + (1 - V_i) \hat{\rho}_{ki,K} \right].$$

It follows that the KNN imputation estimators for TCF_k are

$$\begin{aligned}
\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) &= 1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i < c_1) \left[V_i D_{1i} + (1 - V_i) \hat{\rho}_{1i,K} \right]}{\sum_{i=1}^{n} \left[V_i D_{1i} + (1 - V_i) \hat{\rho}_{1i,K} \right]}, \\
\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1, c_2) &= \frac{\hat{\beta}_{12} - \hat{\beta}_{22}}{\hat{\theta}_2} \\
&= \frac{\sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) \left[V_i D_{2i} + (1 - V_i) \hat{\rho}_{2i,K} \right]}{\sum_{i=1}^{n} \left[V_i D_{2i} + (1 - V_i) \hat{\rho}_{2i,K} \right]}, \qquad (3.1) \\
\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) &= \frac{\hat{\beta}_{23}}{\hat{\theta}_3} = \frac{\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_2) \left[V_i D_{3i} + (1 - V_i) \hat{\rho}_{3i,K} \right]}{\sum_{i=1}^{n} \left[V_i D_{3i} + (1 - V_i) \hat{\rho}_{3i,K} \right]}.
\end{aligned}$$

Note that KNN estimators (3.1) can be seen as nonparametric versions of the MSI estimators (2.2).

3.2. Asymptotic distribution

Let $\rho_k(t, a) = \Pr(D_k = 1 | T = t, A = a)$ and $\pi(t, a) = \Pr(V = 1 | T = t, A = a)$. The KNN imputation estimators of $\mathrm{TCF}_1(c_1)$, $\mathrm{TCF}_2(c_1, c_2)$ and $\mathrm{TCF}_3(c_2)$ are consistent and asymptotically normal. In fact, we have the following theorems.

Theorem 3.1. *Assume the functions $\rho_k(t, a)$ and $\pi(t, a)$ are finite and first-order differentiable. Moreover, assume that the expectation of $1/\pi(T, A)$ exists. Then, for a fixed pair of cut points (c_1, c_2) such that $c_1 < c_2$, the KNN imputation estimators $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1, c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2)$ are consistent.*

Proof: Since the disease status D_k is a Bernoulli random variable, its second-order moment, $\mathbb{E}(D_k^2)$, is finite. According to the first assumption, we can show that the conditional variance of D_k given T and A, $\mathbb{V}\mathrm{ar}(D_k | T = t, A = a)$, is equal to $\rho_k(t, a) \left[1 - \rho_k(t, a) \right]$, which is clearly finite. Thus, by an application of Theorem 1 in [12], the KNN imputation estimators $\hat{\theta}_{k,\mathrm{KNN}}$ are consistent.

Now, observe that, for $j \in \{1,2\}$, $k \in \{1,2,3\}$ and $k \geq j$,

$$\begin{aligned}
\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk} &= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)\left[V_i D_{ki} + (1-V_i)\rho_{ki}\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)(\hat{\rho}_{ki,K} - \rho_{ki}) - \beta_{jk} \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) V_i \left[D_{ki} - \rho_{ki}\right] + \frac{1}{n}\sum_{i=1}^{n}\left[\mathrm{I}(T_i \geq c_j)\rho_{ki} - \beta_{jk}\right] \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)(\hat{\rho}_{ki,K} - \rho_{ki}) \\
&= S_{jk} + R_{jk} + T_{jk}.
\end{aligned}$$

Here, the quantities R_{jk}, S_{jk} and T_{jk} are similar to the quantities R, S and T in the proof of Theorem 2.1 in [4] and of Theorem 1 in [12]. Thus, we have that

$$\begin{aligned}
\sqrt{n}R_{jk} &\xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\right]\right), \\
\sqrt{n}S_{jk} &\xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\pi(T,A)\delta_{jk}^2(T,A)\right]\right),
\end{aligned}$$

where $\delta_{jk}^2(T,A)$ is the conditional variance of $\mathrm{I}(T \geq c_j, D_k = 1)$ given T, A. From proof of Theorem 1 in [12], we also get $T_{jk} = W_{jk} + o_p(n^{-1/2})$, where

$$W_{jk} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j)(1-V_i)\left[\frac{1}{K}\sum_{l=1}^{K}\left(V_{i(l)}D_{ki(l)} - \rho_{ki(l)}\right)\right],$$

with $\mathbb{E}(W_{jk}) = 0$, $\sqrt{n}W_{jk} \xrightarrow{d} \mathcal{N}\left(0, \sigma_{W_{jk}}^2\right)$, and

$$\sigma_{W_{jk}}^2 = \frac{1}{K}\mathbb{E}\left[(1-\pi(T,A))\delta_{jk}^2(T,A)\right] + \mathbb{E}\left[\frac{(1-\pi(T,A))^2\delta_{jk}^2(T,A)}{\pi(T,A)}\right]. \tag{3.2}$$

This leads to the consistency of $\hat{\beta}_{jk,\mathrm{KNN}}$, i.e, $\hat{\beta}_{jk,\mathrm{KNN}} \xrightarrow{p} \beta_{jk}$. It follows that $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) = 1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1}$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1,c_2) = \frac{\hat{\beta}_{12}-\hat{\beta}_{22}}{\hat{\theta}_2}$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) = \frac{\hat{\beta}_{23}}{\hat{\theta}_3}$ are consistent. □

Theorem 3.2. *Assume that the conditions in Theorem 3.1 hold. We get*

$$\sqrt{n}\left[\begin{pmatrix} \widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1) \\ \widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1,c_2) \\ \widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2) \end{pmatrix} - \begin{pmatrix} \mathrm{TCF}_1(c_1) \\ \mathrm{TCF}_2(c_1,c_2) \\ \mathrm{TCF}_3(c_2) \end{pmatrix}\right] \xrightarrow{d} \mathcal{N}(0,\Xi), \tag{3.3}$$

where Ξ is a suitable matrix.

Proof: From proof of Theorem 3.1, we have

$$\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk} = S_{jk} + R_{jk} + W_{jk} + o_p(n^{-1/2}),$$

$\sqrt{n}R_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{V}\mathrm{ar}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\right]\right)$, $\sqrt{n}S_{jk} \xrightarrow{d} \mathcal{N}\left(0, \mathbb{E}\left[\pi(T,A)\delta_{jk}^2(T,A)\right]\right)$ and $\sqrt{n}W_{jk} \xrightarrow{d} \mathcal{N}(0, \sigma_{W_{jk}}^2)$. Moreover, arguments in the proof of Theorem 2.1 in [4] and of Theorem 1 in [12],

allows to state that W_{jk} asymptotically behaves as a sample mean, S_{jk}, R_{jk} and W_{jk} are jointly asymptotically normal, and $\sqrt{n}(\hat{\beta}_{jk,\mathrm{KNN}} - \beta_{jk}) \xrightarrow{d} \mathcal{N}(0, \sigma^2_{jk})$, with $\sigma^2_{jk} = \left[\beta_{jk}\,(1-\beta_{jk}) + \omega^2_{jk}\right]$ and

$$\begin{aligned}\omega^2_{jk} \;=\; & \left(1+\frac{1}{K}\right)\mathbb{E}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))\right] \\ & + \mathbb{E}\left[\mathrm{I}(T \geq c_j)\rho_k(T,A)\frac{(1-\rho_k(T,A))(1-\pi(T,A))^2}{\pi(T,A)}\right].\end{aligned} \tag{3.4}$$

Finally, a direct application of Theorem 1 in [12] gives that $\sqrt{n}(\hat{\theta}_{k,\mathrm{KNN}} - \theta_k)$ converges to a normal random variable with mean 0 and variance $\sigma^2_k = \left[\theta_k(1-\theta_k) + \omega^2_k\right]$, where

$$\begin{aligned}\omega^2_k \;=\; & \left(1+\frac{1}{K}\right)\mathbb{E}\left[\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))\right] \\ & + \mathbb{E}\left[\frac{\rho_k(T,A)(1-\rho_k(T,A))(1-\pi(T,A))^2}{\pi(T,A)}\right].\end{aligned} \tag{3.5}$$

Since $\sqrt{n}\big(\hat{\theta}_{1,\mathrm{KNN}}, \hat{\theta}_{2,\mathrm{KNN}}, \hat{\beta}_{11,\mathrm{KNN}}, \hat{\beta}_{12,\mathrm{KNN}}, \hat{\beta}_{22,\mathrm{KNN}}, \hat{\beta}_{23,\mathrm{KNN}}\big)^\top$ is asymptotically normally distributed with mean $(\theta_1, \theta_2, \beta_{11}, \beta_{12}, \beta_{22}, \beta_{23})^\top$ and suitable covariance matrix Ξ^*, result (3.3) follows by applying the multivariate delta method to

$$h(\hat{\theta}_1, \hat{\theta}_2, \hat{\beta}_{11}, \hat{\beta}_{12}, \hat{\beta}_{22}, \hat{\beta}_{23}) = \left(1 - \frac{\hat{\beta}_{11}}{\hat{\theta}_1}, \frac{(\hat{\beta}_{12} - \hat{\beta}_{22})}{\hat{\theta}_2}, \frac{\hat{\beta}_{23}}{(1-\hat{\theta}_1-\hat{\theta}_2)}\right). \qquad \square$$

Let us denote elements in the asymptotic covariance matrix Ξ as follows

$$\Xi = \begin{pmatrix} \xi_1^2 & \xi_{12} & \xi_{13} \\ \xi_{12} & \xi_2^2 & \xi_{23} \\ \xi_{13} & \xi_{23} & \xi_3^2 \end{pmatrix}.$$

Recall that, from proof of Theorem 3.2, $\sigma^2_k = \left[\theta_k(1-\theta_k) + \omega^2_k\right]$ and $\sigma^2_{jk} = \beta_{jk}(1-\beta_{jk}) + \omega^2_{jk}$, where ω^2_k and ω^2_{jk} are given in (3.5) and (3.4), respectively. In Section S1, Supplementary Material, we show that

$$\begin{aligned}
\xi_1^2 \;=\; & \frac{\beta_{11}^2}{\theta_1^4}\sigma_1^2 + \frac{\sigma_{11}^2}{\theta_1^2} - \frac{\beta_{11}}{\theta_1^3}(\sigma_1^2 + \sigma_{11}^2 - \zeta_{11}^2), \\
\xi_2^2 \;=\; & \sigma_2^2\frac{(\beta_{12}-\beta_{22})^2}{\theta_2^4} + \frac{\lambda^2}{\theta_2^2} - \frac{\beta_{12}-\beta_{22}}{\theta_2^3}(\sigma_{12}^2 - \sigma_{22}^2 - \zeta_{12}^2 + \zeta_{22}^2), \\
\xi_3^2 \;=\; & \frac{\beta_{23}^2\sigma_3^2}{\theta_3^4} + \frac{\sigma_{23}^2}{\theta_3^2} - \frac{\beta_{23}}{\theta_3^3}(\sigma_3^2 + \sigma_{23}^2 - \zeta_{23}^2), \\
\xi_{12} \;=\; & \frac{1}{\theta_1\theta_2}\left[\psi_{1212}^2 + \beta_{11}(\beta_{12}-\beta_{22})\right] - \frac{\beta_{11}}{\theta_1^2\theta_2}\left[\psi_{1212}^2 + \theta_1(\beta_{12}-\beta_{22})\right] \\
& - \frac{\beta_{12}-\beta_{22}}{\theta_2^2\theta_1}\left(\frac{\beta_{11}}{\theta_1}\sigma_{12}^* + \psi_{112}^2 + \theta_2\beta_{11}\right), \\
\xi_{13} \;=\; & \frac{1}{\theta_3}\left[-\frac{\beta_{11}}{\theta_1^2}(\psi_{213}^2 + \theta_1\beta_{23}) + \frac{\psi_{213}^2 + \beta_{11}\beta_{23}}{\theta_1}\right] + \frac{\beta_{23}}{\theta_1\theta_3^2} \\
& \times \left[\frac{\beta_{11}}{\theta_1}\left(\sigma_1^2 + \sigma_{12}^*\right) - \psi_{113}^2 - \theta_3\beta_{11}\right], \\
\xi_{23} \;=\; & \frac{1}{\theta_2\theta_3}\left[-\beta_{23}(\beta_{12}-\beta_{22}) + \frac{\beta_{12}-\beta_{22}}{\theta_2}(\psi_{223}^2 + \theta_2\beta_{23})\right] \\
& + \frac{\beta_{23}}{\theta_2\theta_3^2}\left[\psi_{1223}^2 + \theta_3(\beta_{12}-\beta_{22}) - \frac{\beta_{12}-\beta_{22}}{\theta_2}(\sigma_2^2 + \sigma_{12}^*)\right],
\end{aligned} \tag{3.6}$$

where $\zeta_{jk}^2 = \gamma_{jk}(1-\gamma_{jk}) + \eta_{jk}^2$, $\lambda^2 = (\beta_{12} - \beta_{22})[1 - (\beta_{12} - \beta_{22})] + \omega_{12}^2 - \omega_{22}^2$, $\sigma_{12}^* = -(\theta_1\theta_2 + \psi_{12}^2)$, with $\gamma_{jk} = \Pr(T < c_j, D_k = 1)$ and

$$\begin{aligned}\eta_{jk}^2 &= \frac{K+1}{K}\mathbb{E}\left[\mathrm{I}(T < c_j)\rho_k(T,A)\{1-\rho_k(T,A)\}\{1-\pi(T,A)\}\right] \\ &\quad + \mathbb{E}\left[\mathrm{I}(T < c_j)\rho_k(T,A)\frac{\{1-\rho_k(T,A)\}\{1-\pi(T,A)\}^2}{\pi(T,A)}\right],\end{aligned}$$

$$\begin{aligned}\psi_{12}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\rho_1(T,A)\rho_2(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{1212}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \le T < c_2)\rho_1(T,A)\rho_2(T,A)\right\} \\ &\quad + \mathbb{E}\left\{[1-\pi(T,A)]^2\mathrm{I}(c_1 \le T < c_2)\frac{\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{112}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_2(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_2(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{213}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_2)\rho_1(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_2)\rho_1(T,A)\rho_3(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{113}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_1)\rho_1(T,A)\rho_3(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{223}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(T \ge c_2)\rho_2(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{\frac{[1-\pi(T,A)]^2\mathrm{I}(T \ge c_2)\rho_2(T,A)\rho_3(T,A)}{\pi(T,A)}\right\},\end{aligned}$$

$$\begin{aligned}\psi_{1223}^2 &= \left(1+\frac{1}{K}\right)\mathbb{E}\left\{[1-\pi(T,A)]\mathrm{I}(c_1 \le T < c_2)\rho_2(T,A)\rho_3(T,A)\right\} \\ &\quad + \mathbb{E}\left\{[1-\pi(T,A)]^2\mathrm{I}(c_1 \le T < c_2)\frac{\rho_2(T,A)\rho_3(T,A)}{\pi(T,A)}\right\}.\end{aligned}$$

Therefore, from (3.6), the elements of Ξ depend, among others, on quantities as ω_k^2, ω_{jk}^2, γ_{jk}, η_{jk}^2, ψ_{1212}^2, ψ_{112}^2, ψ_{213}^2, ψ_{12}^2, ψ_{113}^2, ψ_{223}^2 and ψ_{1223}^2. As a consequence, to obtain consistent estimates of the asymptotic variances and covariances, we ultimately need to estimate these quantities.

3.3. Choice of K and of the distance measure

The proposed method is based on nearest-neighbor imputation, which requires the choice of a value for K as well as a distance measure.

In practice, the selection of a suitable distance is typically dictated by features of the data and possible subjective evaluations; thus, a general indication about an adequate choice is difficult to express. In many cases, the simple Euclidean distance may be appropriate. Other times, the researcher may wish to consider specific characteristics of data at hand, and then make a different choice. For example, the diagnostic test result T and the auxiliary covariate A could be heterogeneous with respect to their variances (in particular when the variables are measured on different scales). In this case, the choice of the Mahalanobis distance may be suitable. A further discussion on this topic in the context of medical studies can be found in [8]. Therein, we refer the reader to results relative to numerical datasets.

As for the choice of the size of the neighborhood, [12] argue that nearest-neighbor imputation with a small value of K typically yields negligible bias of the estimators, but a large variance; the opposite happens with a large value of K. The authors suggest that the choice of $K \in \{1, 2\}$ is generally adequate when the aim is to estimate a mean. A similar comment is also raised by [1] and [2], i.e., a small value of K, within the range 1–3, may be a good choice to estimate ROC curves and AUC. However, the authors stress that, in general, the choice of K may depend on the dimension of the feature space, and propose to use cross–validation to find K. Specifically, the authors indicate that a suitable value of the size of neighbor could be found by

$$K^* = \arg\min_K \frac{1}{n_{ver}} \left\| D - \hat{\rho}_K \right\|_1 ,$$

where D is a binary disease status, $\|\cdot\|_1$ denotes L_1 norm for vector and n_{ver} is the number of verified subjects. The formula above can be generalized to our three–class case. In fact, when the disease status has q categories ($q \geq 3$), the difference between $\mathcal{D}$ and $\hat{\rho}_K$ is a $n_{ver} \times (q-1)$ matrix. In such situation, the selection rule could be

$$K^* = \arg\min_K \frac{1}{n_{ver}(q-1)} \left\| \mathcal{D} - \hat{\rho}_K \right\|_{1,1} , \tag{3.7}$$

where $\|\mathcal{A}\|_{1,1}$ denotes $L_{1,1}$ norm of matrix $\mathcal{A}$, i.e.,

$$\|\mathcal{A}\|_{1,1} = \sum_{j=1}^{q-1} \left(\sum_{i=1}^{n_{ver}} |a_{ij}| \right) .$$

4. VARIANCE-COVARIANCE ESTIMATION

Consider first the problem of estimating the variances of $\widehat{\mathrm{TCF}}_{1,\mathrm{KNN}}(c_1)$, $\widehat{\mathrm{TCF}}_{2,\mathrm{KNN}}(c_1, c_2)$ and $\widehat{\mathrm{TCF}}_{3,\mathrm{KNN}}(c_2)$. In a nonparametric framework, quantities as $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2 in Section 3.2 can be estimated by their empirical counterparts, using also the plug–in method. Here, we consider an approach that uses a nearest-neighbor rule to estimate the functions $\rho_k(T, A)$

and the propensity score $\pi(T, A)$, that appear in the expressions of $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2. In particular, for the conditional probabilities of disease, we can use KNN estimates $\tilde{\rho}_{ki} = \hat{\rho}_{ki,\bar{K}}$, where the integer $\bar{K}$ must be greater than one to avoid estimates equal to zero. For the conditional probabilities of verification, we can resort to the KNN procedure proposed in [1], which considers the estimates

$$\tilde{\pi}_i = \frac{1}{K_i^*} \sum_{l=1}^{K_i^*} V_{i(l)},$$

where $\left\{(T_{i(l)}, A_{i(l)}, V_{i(l)}) : l = 1, ..., K_i^*\right\}$ is a set of K_i^* observed triplets and $(T_{i(l)}, A_{i(l)})$ denotes the l-th nearest neighbor to (T_i, A_i) among all (T, A)'s. When V_i equals 0, 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 $V_{i(K_i^*)} = 1$ and $V_i = V_{i(1)} = V_{i(2)} = ... = V_{i(K_i^*-1)} = 0$. In case of $V_i = 1$, K_i^* is such that $V_i = V_{i(1)} = V_{i(2)} = ... = V_{i(K_i^*-1)} = 1$, and $V_{i(K_i^*)} = 0$, i.e., K_i^* is set equal to the rank of the first non–verified nearest neighbor to the unit i. Such a procedure automatically avoids zero values for the $\tilde{\pi}_i$'s.

Then, based on the $\tilde{\rho}_{ki}$'s and $\tilde{\pi}_i$'s, we obtain the estimates

$$\begin{aligned}
\hat{\omega}_k^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right) + \frac{1}{n} \sum_{i=1}^{n} \frac{\tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i}, \\
\hat{\omega}_{jk}^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \mathrm{I}(T_i \geq c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}(T_i \geq c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i}, \\
\hat{\eta}_{jk}^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \mathrm{I}(T_i < c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}(T_i < c_j) \tilde{\rho}_{ki} \left(1 - \tilde{\rho}_{ki}\right) \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i},
\end{aligned}$$

from which, along with $\hat{\theta}_{k,\mathrm{KNN}}$, $\hat{\beta}_{jk,\mathrm{KNN}}$ and

$$\hat{\gamma}_{jk,\mathrm{KNN}} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{I}(T_i < c_j) \left[V_i D_{ki} + (1 - V_i) \hat{\rho}_{ki,K}\right],$$

one derives the estimates of the variances of the proposed KNN imputation estimators.

To obtain estimates of covariances, we need to estimate also the quantities ψ_{1212}^2, ψ_{112}^2, ψ_{213}^2, ψ_{12}^2, ψ_{113}^2, ψ_{223}^2 and ψ_{1223}^2. However, estimates of such quantities are similar to those given above for $\omega_k^2, \omega_{jk}^2$ and η_{jk}^2. For example,

$$\begin{aligned}
\hat{\psi}_{1212}^2 &= \frac{K+1}{nK} \sum_{i=1}^{n} \mathrm{I}(c_1 \leq T_i < c_2) \tilde{\rho}_{1i} \tilde{\rho}_{2i} \left(1 - \tilde{\pi}_i\right) \\
&\quad + \frac{1}{n} \sum_{i=1}^{n} \frac{\mathrm{I}(c_1 \leq T_i < c_2) \tilde{\rho}_{1i} \tilde{\rho}_{2i} \left(1 - \tilde{\pi}_i\right)^2}{\tilde{\pi}_i}.
\end{aligned}$$

Of course, there are other possible approaches to obtain variance and covariance estimates. For instance, one could resort to a standard bootstrap procedure.

5. SIMULATION STUDY

In this section, the ability of KNN method to estimate TCF_1, TCF_2 and TCF_3 is evaluated by using Monte Carlo experiments. We also compare the proposed method with partially parametric approaches, namely, FI, MSI, IPW and SPE approaches. As already mentioned, partially parametric bias-corrected estimators of TCF_1, TCF_2 and TCF_3 require parametric regression models to estimate $\rho_{ki} = \Pr(D_{ki} = 1|T_i, A_i)$, or $\pi_i = \Pr(V_i = 1|T_i, A_i)$, or both. A wrong specification of such models may affect the estimators. Therefore, in the simulation study we consider two scenarios: in the parametric estimation process,

(**i**) the disease model and the verification model are both correctly specified;

(**ii**) the disease model and the verification model are both misspecified.

In both scenarios, we execute 5000 Monte Carlo runs at each setting; we set three sample sizes, i.e., 250, 500 and 1000 in scenario (i) and a sample size of 1000 in scenario (ii).

We consider KNN estimators based on the Euclidean distance, with $K = 1$ and $K = 3$. This in light of the discussion in Section 3.4 and some results of a preliminary simulation study presented in Section S5, Supplementary Material. In such preliminary study, we compared the behavior of the KNN estimators for several choices of the distance measure (Euclidean, Manhattan, Canberra and Mahalanobis) and the size of the neighborhood ($K = 1, 3, 5, 10, 20$).

5.1. Correctly specified parametric models

The true disease is generated by a trinomial random vector (D_1, D_2, D_3), such that D_k is a Bernoulli random variable with success probability θ_k, $k = 1, 2, 3$. We set $\theta_1 = 0.4, \theta_2 = 0.35$ and $\theta_3 = 0.25$. The continuous test result T and a covariate A are generated from the following conditional models

$$T, A|D_k \sim \mathcal{N}_2\left(\mu_k, \Sigma\right), \qquad k = 1, 2, 3,$$

where $\mu_k = (2k, k)^\top$ and

$$\Sigma = \begin{pmatrix} \sigma^2_{T|D} & \sigma_{T,A|D} \\ \sigma_{T,A|D} & \sigma^2_{A|D} \end{pmatrix}.$$

We consider three different values for Σ, specifically

$$\begin{pmatrix} 1.75 & 0.1 \\ 0.1 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 2.5 & 1.5 \\ 1.5 & 2.5 \end{pmatrix}, \qquad \begin{pmatrix} 5.5 & 3 \\ 3 & 2.5 \end{pmatrix},$$

giving rise to a correlation between T and A equal to $0.36, 0.69$ and 0.84, respectively. The verification status V is generated by the following model

$$\mathrm{logit}\left\{\Pr(V = 1|T, A)\right\} = \delta_0 + \delta_1 T + \delta_2 A,$$

where we fix $\delta_0 = 0.5, \delta_1 = -0.3$ and $\delta_2 = 0.75$. This choice corresponds to a verification rate of about 0.65. We consider six pairs of cut points (c_1, c_2), i.e., $(2, 4)$, $(2, 5)$, $(2, 7)$, $(4, 5)$, $(4, 7)$

and $(5, 7)$. Since the conditional distribution of T given D_k is the normal distribution, the true parameters values are

$$\begin{aligned}
\mathrm{TCF}_1(c_1) &= \Phi\left(\frac{c_1 - 2}{\sigma_{T|D}}\right), \\
\mathrm{TCF}_2(c_1, c_2) &= \Phi\left(\frac{c_2 - 4}{\sigma_{T|D}}\right) - \Phi\left(\frac{c_1 - 4}{\sigma_{T|D}}\right), \\
\mathrm{TCF}_3(c_2) &= 1 - \Phi\left(\frac{c_2 - 6}{\sigma_{T|D}}\right),
\end{aligned}$$

where $\Phi(\cdot)$ denotes the cumulative distribution function of the standard normal random variable.

In this set–up, FI, MSI, IPW and SPE estimators are computed under correct working models for both the disease and the verification processes. Therefore, the conditional verification probabilities π_i are estimated from a logistic model for V given T and A with logit link. Under our data–generating process, the true conditional disease model is a multinomial logistic model

$$\Pr(D_k = 1|T, A) = \frac{\exp\left(\tau_{0k} + \tau_{1k}T + \tau_{2k}A\right)}{1 + \exp\left(\tau_{01} + \tau_{11}T + \tau_{21}A\right) + \exp\left(\tau_{02} + \tau_{12}T + \tau_{22}A\right)}$$

for suitable $\tau_{0k}, \tau_{1k}, \tau_{2k}$, where $k = 1, 2$.

Tables 1–3 show Monte Carlo means and standard deviations of the estimators for the three true class factions. Results concern the estimators FI, MSI, IPW, SPE, and the KNN estimator with $K = 1$ and $K = 3$ computed using the Euclidean distance. Also, the estimated standard deviations are shown in the tables. The estimates are obtained by using asymptotic results. To estimate standard deviations of KNN estimators, we use the KNN procedure discussed in Section 4, with $\bar{K} = 2$. Each table refers to a chosen value for Σ. The sample size is 250. The results for sample sizes 500 and 1000 are presented in Section S2 of Supplementary Material.

As expected, the parametric approaches work well when both models for $\rho_k(t, a)$ and $\pi(t, a)$ are correctly specified. FI and MSI estimators seem to be the most efficient ones, whereas the IPW approach seems to provide less powerful estimators, in general. The new proposals (1NN and 3NN estimators) yield also good results, comparable, in terms of bias and standard deviation, to those of the parametric competitors. Moreover, estimators 1NN and 3NN seem to achieve similar performances, and the results about estimated standard deviations of KNN estimators seem to show the effectiveness of the procedure discussed in Section 4.

Finally, some results of simulation experiments performed to explore the effect of a multidimensional vector of auxiliary covariates are given in Section S3, Supplementary Material. A vector A of dimension 3 is employed. The results in Table 7, Supplementary Material, show that KNN estimators still behave satisfactorily.

Table 1: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when $n = 250$ and the first value of Σ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut points $= (2, 4)$				
True	**0.5000**	**0.4347**	**0.9347**						
FI	0.5005	0.4348	0.9344	0.0537	0.0484	0.0269	0.0440	0.0398	0.0500
MSI	0.5005	0.4346	0.9342	0.0550	0.0547	0.0320	0.0465	0.0475	0.0536
IPW	0.4998	0.4349	0.9341	0.0722	0.0727	0.0372	0.0688	0.0702	0.0420
SPE	0.5010	0.4346	0.9344	0.0628	0.0659	0.0364	0.0857	0.0637	0.0363
1NN	**0.4989**	**0.4334**	**0.9331**	0.0592	0.0665	0.0387	0.0555	0.0626	0.0382
3NN	**0.4975**	**0.4325**	**0.9322**	0.0567	0.0617	0.0364	0.0545	0.0608	0.0372
					cut points $= (2, 5)$				
True	**0.5000**	**0.7099**	**0.7752**						
FI	0.5005	0.7111	0.7761	0.0537	0.0461	0.0534	0.0440	0.0400	0.0583
MSI	0.5005	0.7104	0.7756	0.0550	0.0511	0.0566	0.0465	0.0467	0.0626
IPW	0.4998	0.7108	0.7750	0.0722	0.0701	0.0663	0.0688	0.0667	0.0713
SPE	0.5010	0.7106	0.7762	0.0628	0.0619	0.0627	0.0857	0.0604	0.0611
1NN	**0.4989**	**0.7068**	**0.7738**	0.0592	0.0627	0.0652	0.0555	0.0591	0.0625
3NN	**0.4975**	**0.7038**	**0.7714**	0.0567	0.0576	0.0615	0.0545	0.0574	0.0610
					cut points $= (2, 7)$				
True	**0.5000**	**0.9230**	**0.2248**						
FI	0.5005	0.9229	0.2240	0.0537	0.0236	0.0522	0.0440	0.0309	0.0428
MSI	0.5005	0.9231	0.2243	0.0550	0.0285	0.0531	0.0465	0.0353	0.0443
IPW	0.4998	0.9238	0.2222	0.0722	0.0374	0.0765	0.0688	0.0360	0.0728
SPE	0.5010	0.9236	0.2250	0.0628	0.0362	0.0578	0.0857	0.0348	0.0573
1NN	**0.4989**	**0.9201**	**0.2233**	0.0592	0.0372	0.0577	0.0555	0.0366	0.0570
3NN	**0.4975**	**0.9177**	**0.2216**	0.0567	0.0340	0.0558	0.0545	0.0355	0.0563
					cut points $= (4, 5)$				
True	**0.9347**	**0.2752**	**0.7752**						
FI	0.9347	0.2763	0.7761	0.0245	0.0412	0.0534	0.0179	0.0336	0.0583
MSI	0.9348	0.2758	0.7756	0.0271	0.0471	0.0566	0.0220	0.0404	0.0626
IPW	0.9350	0.2758	0.7750	0.0421	0.0693	0.0663	0.0391	0.0651	0.0713
SPE	0.9353	0.2761	0.7762	0.0386	0.0590	0.0627	0.0377	0.0568	0.0611
1NN	**0.9322**	**0.2734**	**0.7738**	0.0374	0.0572	0.0652	0.0342	0.0553	0.0625
3NN	**0.9303**	**0.2712**	**0.7714**	0.0328	0.0526	0.0615	0.0332	0.0538	0.0610
					cut points $= (4, 7)$				
True	**0.9347**	**0.4883**	**0.2248**						
FI	0.9347	0.4881	0.2240	0.0245	0.0541	0.0522	0.0179	0.0444	0.0428
MSI	0.9348	0.4885	0.2243	0.0271	0.0576	0.0531	0.0220	0.0495	0.0443
IPW	0.9350	0.4889	0.2222	0.0421	0.0741	0.0765	0.0391	0.0713	0.0728
SPE	0.9353	0.4890	0.2250	0.0386	0.0674	0.0578	0.0377	0.0646	0.0573
1NN	**0.9322**	**0.4867**	**0.2233**	0.0374	0.0680	0.0577	0.0342	0.0633	0.0570
3NN	**0.9303**	**0.4852**	**0.2216**	0.0328	0.0630	0.0558	0.0332	0.0615	0.0563
					cut points $= (5, 7)$				
True	**0.9883**	**0.2132**	**0.2248**						
FI	0.9879	0.2118	0.2240	0.0075	0.0435	0.0522	0.0055	0.0336	0.0428
MSI	0.9882	0.2127	0.2243	0.0096	0.0467	0.0531	0.0084	0.0388	0.0443
IPW	0.9887	0.2130	0.2222	0.0193	0.0653	0.0765	0.0177	0.0618	0.0728
SPE	0.9888	0.2130	0.2250	0.0191	0.0571	0.0578	0.0184	0.0554	0.0573
1NN	**0.9868**	**0.2133**	**0.2233**	0.0177	0.0567	0.0577	0.0172	0.0532	0.0570
3NN	**0.9860**	**0.2139**	**0.2216**	0.0151	0.0519	0.0558	0.0168	0.0516	0.0563

Table 2: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when $n = 250$ and the second value of Σ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
					cut points $= (2, 4)$				
True	**0.5000**	**0.3970**	**0.8970**						
FI	0.4999	0.3974	0.8973	0.0503	0.0421	0.0362	0.0432	0.0352	0.0466
MSI	0.5000	0.3975	0.8971	0.0521	0.0497	0.0416	0.0461	0.0451	0.0515
IPW	0.4989	0.3990	0.8971	0.0663	0.0685	0.0534	0.0647	0.0681	0.0530
SPE	0.5004	0.3980	0.8976	0.0570	0.0619	0.0516	0.0563	0.0620	0.0493
1NN	**0.4982**	**0.3953**	**0.8976**	0.0587	0.0642	0.0537	0.0561	0.0618	0.0487
3NN	**0.4960**	**0.3933**	**0.8970**	0.0556	0.0595	0.0494	0.0548	0.0600	0.0472
					cut points $= (2, 5)$				
True	**0.5000**	**0.6335**	**0.7365**						
FI	0.4999	0.6337	0.7395	0.0503	0.0436	0.0583	0.0432	0.0379	0.0554
MSI	0.5000	0.6330	0.7385	0.0521	0.0508	0.0613	0.0461	0.0469	0.0612
IPW	0.4989	0.6335	0.7386	0.0663	0.0676	0.0728	0.0647	0.0663	0.0745
SPE	0.5004	0.6333	0.7390	0.0570	0.0622	0.0682	0.0563	0.0612	0.0673
1NN	**0.4982**	**0.6304**	**0.7400**	0.0587	0.0645	0.0721	0.0561	0.0615	0.0672
3NN	**0.4960**	**0.6283**	**0.7396**	0.0556	0.0600	0.0670	0.0548	0.0597	0.0654
					cut points $= (2, 7)$				
True	**0.5000**	**0.8682**	**0.2635**						
FI	0.4999	0.8676	0.2655	0.0503	0.0316	0.0560	0.0432	0.0294	0.0478
MSI	0.5000	0.8678	0.2660	0.0521	0.0374	0.0583	0.0461	0.0364	0.0512
IPW	0.4989	0.8682	0.2669	0.0663	0.0507	0.0698	0.0647	0.0484	0.0692
SPE	0.5004	0.8681	0.2663	0.0570	0.0476	0.0608	0.0563	0.0459	0.0600
1NN	**0.4982**	**0.8672**	**0.2672**	0.0587	0.0495	0.0629	0.0561	0.0458	0.0609
3NN	**0.4960**	**0.8657**	**0.2671**	0.0556	0.0452	0.0610	0.0548	0.0442	0.0601
					cut points $= (4, 5)$				
True	**0.8970**	**0.2365**	**0.7365**						
FI	0.8980	0.2363	0.7395	0.0284	0.0367	0.0583	0.0239	0.0301	0.0554
MSI	0.8976	0.2356	0.7385	0.0318	0.0437	0.0613	0.0292	0.0386	0.0612
IPW	0.8975	0.2345	0.7386	0.0377	0.0594	0.0728	0.0373	0.0578	0.0745
SPE	0.8974	0.2353	0.7390	0.0364	0.0529	0.0682	0.0361	0.0522	0.0673
1NN	**0.8958**	**0.2352**	**0.7400**	0.0388	0.0540	0.0721	0.0373	0.0524	0.0672
3NN	**0.8946**	**0.2350**	**0.7396**	0.0362	0.0502	0.0670	0.0361	0.0510	0.0654
					cut points $= (4, 7)$				
True	**0.8970**	**0.4711**	**0.2635**						
FI	0.8980	0.4703	0.2655	0.0284	0.0512	0.0560	0.0239	0.0413	0.0478
MSI	0.8976	0.4703	0.2660	0.0318	0.0561	0.0583	0.0292	0.0490	0.0512
IPW	0.8975	0.4692	0.2669	0.0377	0.0693	0.0698	0.0373	0.0679	0.0692
SPE	0.8974	0.4701	0.2663	0.0364	0.0638	0.0608	0.0361	0.0629	0.0600
1NN	**0.8958**	**0.4719**	**0.2672**	0.0388	0.0666	0.0629	0.0373	0.0630	0.0609
3NN	**0.8946**	**0.4724**	**0.2671**	0.0362	0.0627	0.0610	0.0361	0.0611	0.0601
					cut points $= (5, 7)$				
True	**0.9711**	**0.2347**	**0.2635**						
FI	0.9710	0.2339	0.2655	0.0124	0.0407	0.0560	0.0104	0.0336	0.0478
MSI	0.9709	0.2348	0.2660	0.0166	0.0461	0.0583	0.0156	0.0412	0.0512
IPW	0.9709	0.2347	0.2669	0.0204	0.0568	0.0698	0.0202	0.0562	0.0692
SPE	0.9709	0.2348	0.2663	0.0202	0.0531	0.0608	0.0199	0.0524	0.0600
1NN	**0.9701**	**0.2368**	**0.2672**	0.0217	0.0549	0.0629	0.0213	0.0533	0.0609
3NN	**0.9695**	**0.2375**	**0.2671**	0.0200	0.0519	0.0610	0.0206	0.0517	0.0601

Table 3: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when $n = 250$ and the third value of Σ is considered. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
cut points $= (2, 4)$									
True	**0.5000**	**0.3031**	**0.8031**						
FI	0.5009	0.3031	0.8047	0.0488	0.0344	0.0495	0.0418	0.0284	0.0467
MSI	0.5005	0.3032	0.8045	0.0515	0.0448	0.0544	0.0460	0.0410	0.0542
IPW	0.5015	0.3030	0.8043	0.0624	0.0632	0.0649	0.0618	0.0620	0.0640
SPE	0.5007	0.3034	0.8043	0.0565	0.0576	0.0628	0.0564	0.0574	0.0614
1NN	**0.4997**	**0.3021**	**0.8047**	0.0592	0.0602	0.0682	0.0571	0.0584	0.0621
3NN	**0.4984**	**0.3018**	**0.8043**	0.0561	0.0565	0.0632	0.0556	0.0566	0.0601
cut points $= (2, 5)$									
True	**0.5000**	**0.4682**	**0.6651**						
FI	0.5009	0.4692	0.6668	0.0488	0.0384	0.0616	0.0418	0.0323	0.0536
MSI	0.5005	0.4687	0.6666	0.0515	0.0495	0.0658	0.0460	0.0455	0.0610
IPW	0.5015	0.4681	0.6670	0.0624	0.0671	0.0753	0.0618	0.0670	0.0743
SPE	0.5007	0.4690	0.6665	0.0565	0.0624	0.0721	0.0564	0.0622	0.0704
1NN	**0.4997**	**0.4676**	**0.6668**	0.0592	0.0661	0.0780	0.0571	0.0634	0.0717
3NN	**0.4984**	**0.4670**	**0.6666**	0.0561	0.0619	0.0729	0.0556	0.0614	0.0695
cut points $= (2, 7)$									
True	**0.5000**	**0.7027**	**0.3349**						
FI	0.5009	0.7030	0.3358	0.0488	0.0375	0.0595	0.0418	0.0318	0.0501
MSI	0.5005	0.7027	0.3360	0.0515	0.0474	0.0637	0.0460	0.0435	0.0563
IPW	0.5015	0.7026	0.3366	0.0624	0.0625	0.0730	0.0618	0.0618	0.0716
SPE	0.5007	0.7032	0.3362	0.0565	0.0591	0.0677	0.0564	0.0583	0.0657
1NN	**0.4997**	**0.7024**	**0.3366**	0.0592	0.0633	0.0712	0.0571	0.0592	0.0675
3NN	**0.4984**	**0.7016**	**0.3362**	0.0561	0.0590	0.0680	0.0556	0.0572	0.0660
cut points $= (4, 5)$									
True	**0.8031**	**0.1651**	**0.6651**						
FI	0.8042	0.1660	0.6668	0.0383	0.0277	0.0616	0.0323	0.0231	0.0536
MSI	0.8037	0.1655	0.6666	0.0415	0.0372	0.0658	0.0380	0.0333	0.0610
IPW	0.8039	0.1651	0.6670	0.0473	0.0503	0.0753	0.0473	0.0493	0.0743
SPE	0.8036	0.1655	0.6665	0.0456	0.0465	0.0721	0.0458	0.0455	0.0704
1NN	**0.8032**	**0.1655**	**0.6668**	0.0487	0.0481	0.0780	0.0472	0.0466	0.0717
3NN	**0.8020**	**0.1651**	**0.6666**	0.0460	0.0450	0.0729	0.0457	0.0451	0.0695
cut points $= (4, 7)$									
True	**0.8031**	**0.3996**	**0.3349**						
FI	0.8042	0.3999	0.3358	0.0383	0.0426	0.0595	0.0323	0.0349	0.0501
MSI	0.8037	0.3995	0.3360	0.0415	0.0522	0.0637	0.0380	0.0463	0.0563
IPW	0.8039	0.3996	0.3366	0.0473	0.0658	0.0730	0.0473	0.0645	0.0716
SPE	0.8036	0.3998	0.3362	0.0456	0.0618	0.0677	0.0458	0.0606	0.0657
1NN	**0.8032**	**0.4003**	**0.3366**	0.0487	0.0660	0.0712	0.0472	0.0619	0.0675
3NN	**0.8020**	**0.3998**	**0.3362**	0.0460	0.0617	0.0680	0.0457	0.0600	0.0660
cut points $= (5, 7)$									
True	**0.8996**	**0.2345**	**0.3349**						
FI	0.9003	0.2338	0.3358	0.0266	0.0351	0.0595	0.0224	0.0292	0.0501
MSI	0.9004	0.2340	0.3360	0.0308	0.0443	0.0637	0.0285	0.0398	0.0563
IPW	0.9005	0.2345	0.3366	0.0355	0.0555	0.0730	0.0353	0.0550	0.0716
SPE	0.9004	0.2342	0.3362	0.0349	0.0523	0.0677	0.0346	0.0517	0.0657
1NN	**0.9000**	**0.2348**	**0.3366**	0.0373	0.0556	0.0712	0.0361	0.0531	0.0675
3NN	**0.8992**	**0.2346**	**0.3362**	0.0349	0.0520	0.0680	0.0349	0.0515	0.0660

5.2. Misspecified models

We start from two independent random variables $Z_1 \sim \mathcal{N}(0, 0.5)$ and $Z_2 \sim \mathcal{N}(0, 0.5)$. The true conditional disease is generated by a trinomial random vector (D_1, D_2, D_3) such that

$$D_1 = \begin{cases} 1 \text{ if } Z_1 + Z_2 \leq h_1 \\ 0 \text{ otherwise} \end{cases}, \qquad D_2 = \begin{cases} 1 \text{ if } h_1 < Z_1 + Z_2 \leq h_2 \\ 0 \text{ otherwise} \end{cases},$$

and

$$D_3 = \begin{cases} 1 \text{ if } Z_1 + Z_2 > h_2 \\ 0 \text{ otherwise} \end{cases}.$$

Here, h_1 and h_2 are two thresholds. We choose h_1 and h_2 to make $\theta_1 = 0.4$ and $\theta_3 = 0.25$. The continuous test results T and the covariate A are generated to be related to $\mathcal{D}$ through Z_1 and Z_2. More precisely,

$$T = \alpha(Z_1 + Z_2) + \varepsilon_1, \qquad A = Z_1 + Z_2 + \varepsilon_2,$$

where ε_1 and ε_2 are two independent normal random variables with mean 0 and the common variance 0.25. We choose $\alpha = 0.5$. The verification status V is simulated by the following logistic model

$$\text{logit}\left\{\Pr(V = 1|T, A)\right\} = -1.5 - 0.35T - 1.5A.$$

Under this model, the verification rate is roughly 0.276. For the cut-point, we consider six pairs (c_1, c_2), i.e., $(-1.0, -0.5)$, $(-1.0, 0.7)$, $(-1.0, 1.3)$, $(-0.5, 0.7)$, $(-0.5, 1.3)$ and $(0.7, 1.3)$. Within this set–up, we determine the true values of TCF's as follows:

$$\begin{aligned}
\text{TCF}_1(c_1) &= \frac{1}{\Phi(h_1)} \int_{-\infty}^{h_1} \Phi\left(\frac{c_1 - \alpha z}{\sqrt{0.25}}\right) \phi(z)\mathrm{d}z, \\
\text{TCF}_2(c_1, c_2) &= \frac{1}{\Phi(h_2) - \Phi(h_1)} \int_{h_1}^{h_2} \left[\Phi\left(\frac{c_2 - \alpha z}{\sqrt{0.25}}\right) - \Phi\left(\frac{c_1 - \alpha z}{\sqrt{0.25}}\right)\right] \phi(z)\mathrm{d}z, \\
\text{TCF}_3(c_2) &= 1 - \frac{1}{1 - \Phi(h_2)} \int_{h_2}^{\infty} \Phi\left(\frac{c_2 - \alpha z}{\sqrt{0.25}}\right) \phi(z)\mathrm{d}z,
\end{aligned}$$

where $\phi(\cdot)$ denotes the density function of the standard normal random variable.

The aim in this scenario is to compare FI, MSI, IPW, SPE and KNN estimators when both the estimates for $\hat{\pi}_i$ and $\hat{\rho}_{ki}$ in the parametric approach are inconsistent. Therefore, $\hat{\rho}_{ki}$ is obtained from a multinomial logistic regression model with $\mathcal{D} = (D_1, D_2, D_3)$ as the response and T as predictor. To estimate π_i, we use a generalized linear model for V given T and $A^{2/3}$ with logit link. Clearly, the two fitted models are misspecified. The KNN estimators are obtained by using $K = 1$ and $K = 3$ and the Euclidean distance. Again, we use $\bar{K} = 2$ in the KNN procedure to estimate standard deviations of KNN estimators. As a large sample size is required to guarantee that FI, MSI, IPW, SPE and KNN estimators reach a substantial stability, we set $n = 1000$. For KNN estimators, results based on smaller sample sizes are reported in Section S4, Supplementary Material.

Table 4 presents Monte Carlo means and standard deviations (across 5000 replications) for the estimators of the true class fractions, TCF_1, TCF_2 and TCF_3. The table also gives the means of the estimated standard deviations (of the estimators), based on the asymptotic theory.

Table 4: Monte Carlo means, Monte Carlo standard deviations and estimated standard deviations of the estimators for the true class fractions, when both models for $\rho_k(t,a)$ and $\pi(t,a)$ are misspecified and the sample size $n = 1000$. "True" denotes the true parameter value.

	TCF_1	TCF_2	TCF_3	$MC.sd_1$	$MC.sd_2$	$MC.sd_3$	$asy.sd_1$	$asy.sd_2$	$asy.sd_3$
cut points $= (-1.0, -0.5)$									
True	**0.1812**	**0.1070**	**0.9817**						
FI	0.1290	0.0588	0.9888	0.0153	0.0133	0.0118	0.0170	0.0126	0.0423
MSI	0.1299	0.0592	0.9895	0.0154	0.0153	0.0131	0.0171	0.0144	0.0427
IPW	0.1231	0.0576	0.9889	0.0178	0.0211	0.0208	0.0174	0.0201	0.2878
SPE	0.1407	0.0649	0.9877	0.0173	0.0216	0.0231	0.0171	0.0207	0.0125
1NN	**0.1809**	**0.1036**	**0.9817**	0.0224	0.0304	0.0255	0.0210	0.0257	0.0180
3NN	**0.1795**	**0.0991**	**0.9814**	0.0214	0.0258	0.0197	0.0207	0.0240	0.0190
cut points $= (-1.0, 0.7)$									
True	**0.1812**	**0.8609**	**0.4469**						
FI	0.1290	0.7399	0.5850	0.0153	0.0447	0.1002	0.0170	0.0403	0.0919
MSI	0.1299	0.7423	0.5841	0.0154	0.0453	0.1008	0.0171	0.0408	0.0926
IPW	0.1231	0.7690	0.5004	0.0178	0.0902	0.2049	0.0174	0.0824	0.1844
SPE	0.1407	0.7635	0.5350	0.0173	0.0702	0.2682	0.0171	0.0646	0.2171
1NN	**0.1809**	**0.8452**	**0.4406**	0.0224	0.0622	0.1114	0.0210	0.0503	0.0895
3NN	**0.1795**	**0.8285**	**0.4339**	0.0214	0.0521	0.0882	0.0207	0.0479	0.0929
cut points $= (-1.0, 1.3)$									
True	**0.1812**	**0.9732**	**0.1171**						
FI	0.1290	0.9499	0.1900	0.0153	0.0179	0.0550	0.0170	0.0203	0.0440
MSI	0.1299	0.9516	0.1902	0.0154	0.0184	0.0552	0.0171	0.0206	0.0442
IPW	0.1231	0.9645	0.1294	0.0178	0.0519	0.1795	0.0174	0.0268	0.0898
SPE	0.1407	0.9567	0.1760	0.0173	0.0425	0.3383	0.0171	0.0311	0.2127
1NN	**0.1809**	**0.9656**	**0.1124**	0.0224	0.0218	0.0448	0.0210	0.0272	0.0544
3NN	**0.1795**	**0.9604**	**0.1086**	0.0214	0.0172	0.0338	0.0207	0.0262	0.0567
cut points $= (-0.5, 0.7)$									
True	**0.4796**	**0.7539**	**0.4469**						
FI	0.3715	0.6811	0.5850	0.0270	0.0400	0.1002	0.0244	0.0353	0.0919
MSI	0.3723	0.6831	0.5841	0.0271	0.0409	0.1008	0.0246	0.0361	0.0926
IPW	0.3547	0.7114	0.5004	0.0325	0.0883	0.2049	0.0321	0.0815	0.1844
SPE	0.3949	0.6986	0.5350	0.0318	0.0687	0.2682	0.0312	0.0637	0.2171
1NN	**0.4783**	**0.7416**	**0.4406**	0.0361	0.0610	0.1114	0.0310	0.0526	0.0895
3NN	**0.4756**	**0.7294**	**0.4339**	0.0341	0.0499	0.0882	0.0303	0.0500	0.0929
cut points $= (-0.5, 1.3)$									
True	**0.4796**	**0.8661**	**0.1171**						
FI	0.3715	0.8910	0.1900	0.0270	0.0202	0.0550	0.0244	0.0218	0.0440
MSI	0.3723	0.8924	0.1902	0.0271	0.0211	0.0552	0.0246	0.0226	0.0442
IPW	0.3547	0.9068	0.1294	0.0325	0.0535	0.1795	0.0321	0.0384	0.0898
SPE	0.3949	0.8918	0.1760	0.0318	0.0451	0.3383	0.0312	0.0368	0.2127
1NN	**0.4783**	**0.8620**	**0.1124**	0.0361	0.0349	0.0448	0.0310	0.0373	0.0544
3NN	**0.4756**	**0.8613**	**0.1086**	0.0341	0.0285	0.0338	0.0303	0.0355	0.0567
cut points $= (0.7, 1.3)$									
True	**0.9836**	**0.1122**	**0.1171**						
FI	0.9618	0.2099	0.1900	0.0122	0.0317	0.0550	0.0114	0.0263	0.0440
MSI	0.9613	0.2093	0.1902	0.0125	0.0320	0.0552	0.0116	0.0265	0.0442
IPW	0.9548	0.1955	0.1294	0.0339	0.0831	0.1795	0.0278	0.0764	0.0898
SPE	0.9582	0.1932	0.1760	0.0332	0.0618	0.3383	0.0290	0.0577	0.2127
1NN	**0.9821**	**0.1204**	**0.1124**	0.0144	0.0494	0.0448	0.0109	0.0449	0.0544
3NN	**0.9804**	**0.1319**	**0.1086**	0.0138	0.0404	0.0338	0.0108	0.0429	0.0567

The table clearly shows limitations of the (partially) parametric approaches in case of misspecified models for $\Pr(D_k = 1|T, A)$ and $\Pr(V = 1|T, A)$. More precisely, in term of bias, the FI, MSI, IPW and SPE approaches perform almost always poorly, with high distortion in almost all cases. As we mentioned in Section 2, the SPE estimators could fall outside the interval $(0, 1)$. In our simulations, in the worst case, the estimator $\widehat{\mathrm{TCF}}_{3,\mathrm{SPE}}(-1.0, -0.5)$ gives rise to 20% of the values greater than 1. Moreover, the Monte Carlo standard deviations shown in the table indicate that the SPE approach might yield unstable estimates. Finally, the misspecification also has a clear effect on the estimated standard deviations of the estimators. On the other side, the estimators 1NN and 3NN seem to perform well in terms of both bias and standard deviation. In fact, KNN estimators yield estimated values that are near to the true values. In addition, we observe that the estimator 3NN has larger bias than 1NN, but with slightly less variance.

6. AN ILLUSTRATION

We use data on epithelial ovarian cancer (EOC) extracted from the Pre-PLCO Phase II Dataset from the SPORE/Early Detection Network/Prostate, Lung, Colon and Ovarian Cancer Ovarian Validation Study. [1]

As in [16], we consider the following three classes of EOC, i.e., benign disease, early stage (I and II) and late stage (III and IV) cancer, and 12 of the 59 available biomarkers, i.e. CA125, CA153, CA72–4, Kallikrein 6 (KLK6), HE4, Chitinase (YKL40) and immune costimulatory protein–B7H4 (DD–0110), Insulin–like growth factor 2 (IGF2), Soluble mesothelin-related protein (SMRP), Spondin–2 (DD–P108), Decoy Receptor 3 (DcR3; DD–C248) and Macrophage inhibitory cytokine 1 (DD–X065). In addition, age of patients is also considered.

After cleaning for missing data, we are left 134 patients with benign disease, 67 early stage samples and 77 late stage samples. As a preliminary step of our analysis we ranked the 12 markers according to value of VUS, estimated on the complete data. The observed ordering, consistent with medical knowledge, led us to select CA125 as the test T to be used to illustrate our method.

To mimic verification bias, a subset of the complete dataset is constructed using the test T and a vector $A = (A_1, A_2)$ of two covariates, namely the marker CA153 (A_1) and age (A_2). Reasons for using CA153 as a covariate come from the medical literature that suggests that the concomitant measurement of CA153 with CA125 could be advantageous in the pre-operative discrimination of benign and malignant ovarian tumors. In this subset, T and A are known for all samples (patients), but the true status (benign, early stage or late stage) is available only for some samples, that we select according to the following mechanism. We select all samples having a value for T, A_1 and A_2 above their respective medians, i.e. 0.87, 0.30 and 45; as for the others, we apply the following selection process

$$\Pr(V = 1|T, A) = 0.05 + 0.35\mathrm{I}(T > 0.87) + 0.25\mathrm{I}(A_1 > 0.30) + 0.35\mathrm{I}(A_2 > 45),$$

leading to a marginal probability of selection equal to 0.634.

[1] The study protocol and data are publicly available at the address: https://edrn.nci.nih.gov/protocols/119-spore-edrn-pre-plco-ovarian-phase-ii-validation.

Since the test T and the covariates A_1, A_2 are heterogeneous with respect to their variances, the Mahalanobis distance is used for KNN estimators. Based on the discussion in Section 3.4, we use the selection rule (3.7) to find the size K of the neighborhood. This leads to the choice of $K = 1$ for our data. In addition, we also employ $K = 3$ for the sake of comparison with 1NN result, and produce the estimate of the ROC surface based on full data (Full estimate), displayed in Figure 1.

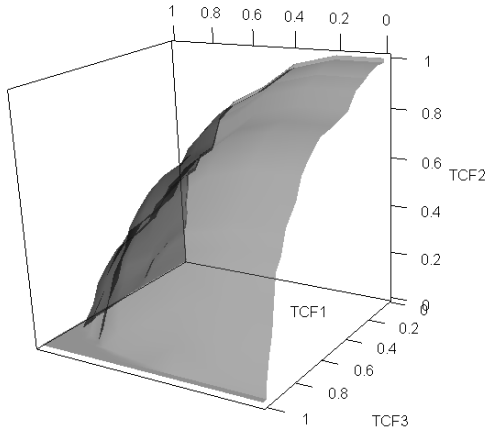
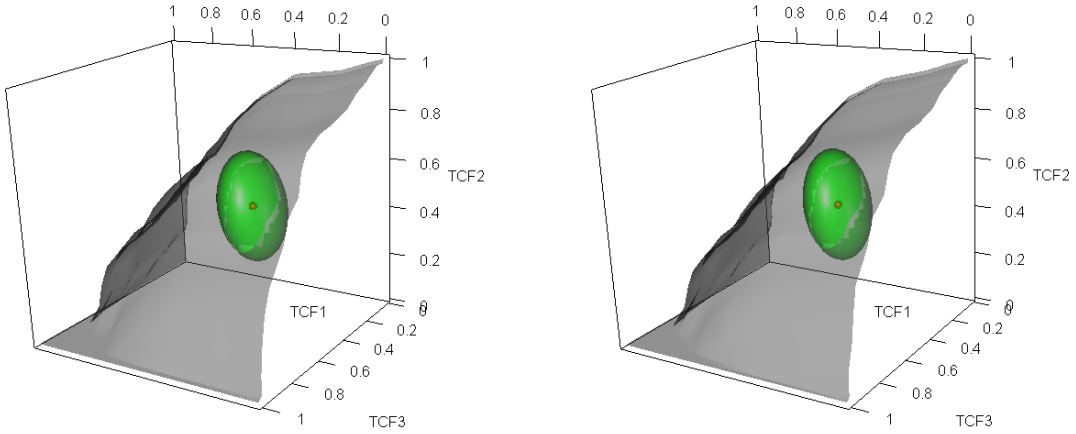


Figure 1: Estimated ROC surface for CA125, based on full data.

Figure 2 shows the 1NN and 3NN estimated ROC surfaces for the test T (CA125).



(a) 1NN (b) 3NN

Figure 2: Bias–corrected estimated ROC surfaces for CA125, based on incomplete data.

In this figure, we also give the 95% ellipsoidal confidence regions (green color) for (TCF_1, $\mathrm{TCF}_2, \mathrm{TCF}_3$) at cut points $(-0.56, 2.31)$. These regions are built using the asymptotic normality of the estimators. Compared with the Full estimate, KNN bias-corrected method

proposed in the paper appears to well behave, yielding reasonable estimates of the ROC surface with incomplete data. A closer inspection to the behavior at some chosen points can be taken by looking at Table 5.

Table 5: Comparison between Full and KNN estimates of the true class fractions for CA125, for some values of c_1 and c_2.

	Full			1NN			3NN		
(c_1, c_2)	TCF_1	TCF_2	TCF_3	TCF_1	TCF_2	TCF_3	TCF_1	TCF_2	TCF_3
(0, 0.5)	0.500	0.104	0.922	0.516	0.171	0.938	0.497	0.170	0.933
(0, 1)	0.500	0.254	0.883	0.516	0.271	0.838	0.497	0.275	0.858
(0, 2.6)	0.500	0.567	0.688	0.516	0.557	0.663	0.497	0.550	0.667
(0, 3)	0.500	0.612	0.623	0.516	0.614	0.612	0.497	0.605	0.617
(0, 4)	0.500	0.731	0.325	0.516	0.714	0.312	0.497	0.710	0.317
(0.4, 0.5)	0.694	0.030	0.922	0.688	0.043	0.938	0.670	0.040	0.933
(0.4, 1)	0.694	0.179	0.883	0.688	0.143	0.838	0.670	0.145	0.858
(0.4, 2.6)	0.694	0.493	0.688	0.688	0.429	0.663	0.670	0.420	0.667
(0.4, 3)	0.694	0.537	0.623	0.688	0.486	0.612	0.670	0.475	0.617
(0.4, 4)	0.694	0.657	0.325	0.688	0.586	0.312	0.670	0.580	0.317
(1, 2.6)	0.813	0.313	0.688	0.789	0.286	0.663	0.787	0.275	0.667
(1, 3)	0.813	0.358	0.623	0.789	0.343	0.612	0.787	0.330	0.617
(1, 4)	0.813	0.478	0.325	0.789	0.443	0.312	0.787	0.435	0.317
(2, 2.6)	0.955	0.149	0.688	0.945	0.143	0.663	0.942	0.130	0.667
(2, 3)	0.955	0.194	0.623	0.945	0.200	0.612	0.942	0.185	0.617
(2, 4)	0.955	0.313	0.325	0.945	0.300	0.312	0.942	0.290	0.317
(3.5, 4)	0.993	0.045	0.325	0.992	0.043	0.312	0.990	0.045	0.317

7. CONCLUSIONS

A general suitable strategy for reducing the effects of model misspecification in statistical inference is to resort on fully nonparametric methods. This paper proposes a nonparametric estimator of the ROC surface of a continuous diagnostic test. The estimator is based on nearest-neighbor imputation and works under MAR assumption. It represents an alternative to (partially) parametric estimators discussed in [16]. Our simulation results and the presented illustrative example show usefulness of the proposal.

Generally speaking, performances of our estimator depend on various intrinsic factors, and on some user-defined choices. Among intrinsic factors, we mention the unknown values of parameters TCF_1, TCF_2 and TCF_3 to be estimated, the rate of verified units in the sample at hand, and the nature of the unknown processes generating the observations. In particular, extreme values of the true class fractions, i.e. values close to 0 or 1, are difficult to estimate in an accurate way, especially when sample data are characterized by a low verification rate, which limits the amount of information available. On the basis of discussions in Section 3.3 (and in the last part of this section) and of simulation results in Section 5 (and in Supplementary Material), we offer some recommendations for tackling the user-defined choices. More precisely, we recommend: (a) to use the Euclidean distance, as the first choice,

and the Mahalanobis distance in case of heterogeneity among variables; (b) to keep small, from 1 to 3, say, the number of neighbors K. Our simulation results show satisfactory performances of the KNN estimator of the ROC surface when about 70 verified observations are present in the sample.

As in [1], a simple extension of our estimator, that could be used when categorical auxiliary variables are also available, is possible. Without loss of generality, we suppose that a single factor C, with m levels, is observed together with T and A. We also assume that C may be associated with both $\mathcal{D}$ 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, for example, if the MAR assumption and first-order differentiability of the functions $\rho_k(t,a)$ and $\pi(t,a)$ hold in each stratum, a consistent and asymptotically normally distributed estimator of TCF_1 is

$$\widehat{\mathrm{TCF}}^{S}_{1,\mathrm{KNN}}(c_1) = \frac{1}{n}\sum_{j=1}^{m} n_j \widehat{\mathrm{TCF}}^{cond}_{1j,\mathrm{KNN}}(c_1),$$

where n_j denotes the size of the j-th stratum and the quantity $\widehat{\mathrm{TCF}}^{cond}_{1j,\mathrm{KNN}}(c_1)$ denotes the KNN estimator of the conditional TCF_1, i.e., the KNN estimator in (3.1) 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.

In our approach, the KNN method is used to estimate the probabilities $\rho_k(t,a)$ for non–verified subjects. A referee pointed out that KNN estimators might suffer from boundary effects, i.e., increases in bias when estimates are computed near the boundary of the support of the covariates. Indeed, near the boundaries, any smoothing method is less accurate, as fewer observations can be averaged, so that bias of estimators can be affected. In contrast to other nonparametric regression methods, however, KNN estimators always involve the same number of observations. Boundary effects, therefore, act on neighborhoods' sizes more than on the number of observations involved in the local fitting. For this reason, a prominent source of bias of KNN estimators is the shape at the boundary of the functions to be estimated. Steeper functions are more likely associated to a larger bias, an aspect pointing to small values of K as good choices to limit boundary effects. Moreover, it is worth noting that in the domain of our interest, i.e., evaluation of diagnostic tests, is hard to deal with test and covariate values close to the boundary of their support. More likely, one faces sparsity of data in some regions of the features space and, therefore, one has to deal with situations in which, for a fixed sample size, information brought by data on those regions is structurally low. This aspect also impacts on the neighborhoods' sizes, and probably amounts to a primary source of bias in our application contest. This remark is supported by results of some simulations that we carried out to evaluate possible bias due to boundary effects and/or sparsity of data (see Section S5, Supplementary Material). Overall, simulation results seem to show that the bias, when present, is driven more by sparsity of data issues than by boundary effects and that KNN estimators have their poorest performances on largest values of K, regardless of the position of points in the domain.

ACKNOWLEDGMENTS

This work has been supported by the grant number BIRD169208 from University of Padova, Italy. We also acknowledge the valuable suggestions from the Associated Editor and two anonymous Referees, who greatly contributed to improve presentation of the contents.

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.

[2] Adimari, G. and Chiogna, M. (2017). Nonparametric verification bias-corrected inference for the area under the ROC curve of a continuous-scale diagnostic test, *Statistics and Its Interface*, **10**, 629–641.

[3] 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**, 173–290.

[4] Cheng, P.E. (1994). Nonparametric estimation of mean functionals with data missing at random, *Journal of the American Statistical Association*, **89**, 81–87.

[5] Chi, Y.Y. and Zhou, X.H. (2008). Receiver operating characteristic surfaces in the presence of verification bias, *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, **57**, 1–23.

[6] Dreiseitl, S.; Ohno-Machado, L. and Binder, M. (2000). Comparing three-class diagnostic tests by three-way ROC analysis, *Medical Decision Making*, **20**(3), 323–331.

[7] He, H. and McDermott, M.P. (2012). A robust method using propensity score stratification for correcting verification bias for binary tests, *Biostatistics*, **13**, 32–47.

[8] Hu, L.-Y.; Huang, M.-W.; Ke, S.-W. and Tsai, C.-F. (2016). The distance function effect on k-nearest neighbor classification for medical datasets, *SpringerPlus*, **5**(1), 1304.

[9] Little, R.J.A. and Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, Wiley, New York.

[10] Nakas, C.T. (2014). Developments in ROC surface analysis and assessment of diagnostic markers in three-class classification problems, *REVSTAT – Statistical Journal*, **12**, 43–65.

[11] Nakas, C.T. and Yiannoutsos, C.Y. (2004). Ordered multiple-class ROC analysis with continuous measurements, *Statistics in Medicine*, **23**, 3437–3449.

[12] Ning, J. and Cheng, P.E. (2012). A comparison study of nonparametric imputation methods, *Statistics and Computing*, **22**, 273–285.

[13] Pepe, M.S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction*, Oxford University Press.

[14] 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.

[15] Scurfield, B.K. (1996). Multiple-event forced-choice tasks in the theory of signal detectability, *Journal of Mathematical Psychology*, **40**, 253–269.

[16] To Duc, K.; Chiogna, M. and Adimari, G. (2016). Bias-corrected methods for estimating the receiver operating characteristic surface of continuous diagnostic tests, *Electronic Journal of Statistics*, **10**(2), 3063–3113.

[17] Zhou, X.H.; Obuchowski, N.A. and McClish, D.K. (2002). *Statistical Methods in Diagnostic Medicine*, Wiley and Sons, New York.