Received: 27 July 2020 | Revised: 12 July 2021 | Accepted: 19 July 2021

DOI: 10.1002/sim.9153

RESEARCH ARTICLE

Statistics in Medicine WILEY

Robust and flexible inference for the covariate-specific receiver operating characteristic curve

Vanda Inácio[1] | **Vanda M. Lourenço**[2] | **Miguel de Carvalho**[1] | **Richard A. Parker**[3] | **Vincent Gnanapragasam**[4,5,6]

[1]School of Mathematics, University of Edinburgh, Edinburgh, UK

[2]Department of Mathematics and CMA, NOVA School of Sciences and Technology, NOVA University of Lisbon, Caparica, Portugal

[3]Edinburgh Clinical Trials Unit, Usher Institute, University of Edinburgh, Edinburgh, UK

[4]Cambridge Urology Translational Research and Clinical Trials Office, Cambridge, UK

[5]Academic Urology Group, Department of Surgery, University of Cambridge, Cambridge, UK

[6]Department of Urology, Cambridge University Hospitals Trust, Cambridge, UK

Correspondence
Vanda Inácio, School of Mathematics, University of Edinburgh, The King's Buildings, JCMB, Edinburgh EH9 3FD, Scotland, UK.
Email: vanda.inacio@ed.ac.uk

Funding information
Fundação para a Ciência e Tecnologia, Grant/Award Numbers: PTDC/MAT-STA/28649/2017, UID/MAT/00006/2020, UIDB/00297/2020, SFRH/BSAB/142919/2018; Erasmus, Grant/Award Numbers: 29191/002/2017/STT, 29191/036/2018/STT, 032/2020/SAM/FCT

Diagnostic tests are of critical importance in health care and medical research. Motivated by the impact that atypical and outlying test outcomes might have on the assessment of the discriminatory ability of a diagnostic test, we develop a robust and flexible model for conducting inference about the covariate-specific receiver operating characteristic (ROC) curve that safeguards against outlying test results while also accommodating for possible nonlinear effects of the covariates. Specifically, we postulate a location-scale regression model for the test outcomes in both the diseased and nondiseased populations, combining additive regression B-splines and M-estimation for the regression function, while the distribution of the error term is estimated via a weighted empirical distribution function of the standardized residuals. The results of the simulation study show that our approach successfully recovers the true covariate-specific area under the ROC curve on a variety of conceivable test outcomes contamination scenarios. Our method is applied to a dataset derived from a prostate cancer study where we seek to assess the ability of the Prostate Health Index to discriminate between men with and without Gleason 7 or above prostate cancer, and if and how such discriminatory capacity changes with age.

KEYWORDS
additive regression B-splines, covariate-adjustment, diagnostic test, M-estimation, outliers, receiver operating characteristic curve

1 | INTRODUCTION

The evaluation of the performance of a medical test for screening and diagnosing disease is an important step toward advancing health in individuals and communities. The major goal of a diagnostic test is to distinguish diseased from nondiseased individuals or, more generally, to distinguish between different disease stages. Before the widespread use of a test, its ability to discriminate between the different disease states must be rigorously vetted. Note that here we

use the term "diagnostic test," or sometimes simply "test," to broadly encompass any continuous classifier, which may include a single biological marker or a composite score resulting from the combination of multiple biomarkers. We further note that we will be assuming the existence of a so-called gold standard test, that is, a perfect test that correctly classifies all individuals as being diseased or nondiseased. Compared to the diagnosis made by the gold standard test, the goal is to assess how well the candidate test, which is possibly less invasive and/or costly, performs. The receiver operating characteristic (ROC) curve is the most popular graphical tool used for evaluating the discriminatory ability of continuous-outcome tests. The ROC curve is a plot of the false positive fraction (probability that a nondiseased subject tests positive) against the true positive fraction (probability that a diseased subject tests positive) for all possible threshold values that can be used to convert continuous test outcomes into binary ones. Further background on ROC curves is provided in Section 2.

It has been recognized that the performance of a test may be affected by covariates, such as age and/or gender and, in such situations, ignoring covariate information might result in erroneous conclusions about a test's accuracy. The full understanding of how covariates impact a test's performance is thus of paramount importance in order to determine the optimal and suboptimal populations, as defined by the covariate values, in which to perform the tests. The covariate-specific or conditional ROC curve, which is an ROC curve that conditions on a specific covariate value, arises as the natural tool to use in this context. For a recent overview of available ROC regression methods, we refer to Inácio et al.[1]

Motivated by the fact that atypical/outlying test outcomes (due, for instance, to experimental, biological, or coding errors) may put at risk the reliability of the inferences about the test's accuracy, we develop a robust additive regression B-splines modeling framework for conducting inference about the covariate-specific ROC curve that mitigates the impact that outliers can have on inferences, while simultaneously allowing for nonlinear effects of the covariates. Here and below, by an outlier or atypical test outcome we mean an outcome that is clearly separated from the majority or bulk of the test outcomes, or that in some way deviates from the general patterns present in the test results.[2(p124)] Our estimation method for the covariate-specific ROC curve is similar in spirit to those developed by Pepe,[3] González-Manteiga et al,[4] Rodríguez-Álvarez et al,[5] and Rodríguez and Martínez,[6] which postulates a location-scale regression model for the test outcomes in both the diseased and nondiseased populations (termed in the literature as "induced" approach). Yet, unlike previous approaches: (i) our specification for the regression function relies on an additive regression B-splines formulation, with M-estimation used for the regression coefficients, hence safeguarding against outlying test outcomes, and (ii) the distribution of the regression errors is modeled via a weighted empirical distribution function of the standardized residuals, therefore downweighting the influence of outliers when estimating the covariate-specific ROC curve and its associated summary indices. These features result in a widely applicable approach that can be used for many populations and for a large number of diseases and continuous diagnostic tests. In addition, from a computational perspective, our method is extremely fast and can be easily implemented in any software package. We acknowledge that the approaches of González-Manteiga et al,[4] Rodríguez-Álvarez et al,[5] and Rodríguez and Martínez[6] also allow for nonlinear effects of the covariates on the mean (and, unlike ours, also on the variance) function but, unlike our proposed approach, they do it through the use of kernel methods (the former two approaches) and Gaussian processes (the latter approach).

The remainder of this article is organized as follows. In Section 2, we introduce our modeling approach to conduct inference about the covariate-specific ROC curve. The performance of our method is validated in Section 3 using simulated data under different test results' and not results" contamination scenarios. In Section 4, our approach is applied to assess the age-specific accuracy of the Prostate Health Index (PHI) as a biomarker for prostate cancer. Concluding remarks are offered in Section 5.

2 | ROBUST AND FLEXIBLE INFERENCE FOR THE COVARIATE-SPECIFIC ROC CURVE

2.1 | Preliminaries

We start with some background on ROC curves. Let Y be the continuous random variable denoting the outcome of the diagnostic test and D the binary variable indicating the presence ($D = 1$) or absence ($D = 0$) of disease. Throughout, we use the subscripts D and $\overline{D}$ to denote quantities conditional on $D = 1$ and $D = 0$, respectively. For example, Y_D and $Y_{\overline{D}}$ denote the test outcomes in the diseased and nondiseased populations, with cumulative distribution functions given by F_D and $F_{\overline{D}}$, respectively. Further, let c be the threshold value used for defining a positive test result. Without loss of generality,

we proceed with the assumption that larger values of Y are more indicative of disease; that is, a subject is diagnosed as diseased when his/her test outcome is equal or greater than c, $Y \geq c$, and he or she is diagnosed as nondiseased when the outcome is below c, $Y < c$. Hence, for each possible threshold c, the true positive fraction (TPF) and false positive fraction (FPF) corresponding to such decision criterion are

$$\begin{aligned} \text{TPF}(c) &= \Pr(Y \geq c \,|\, D = 1) = \Pr(Y_D \geq c) = 1 - F_D(c), \\ \text{FPF}(c) &= \Pr(Y \geq c \,|\, D = 0) = \Pr(Y_{\overline{D}} \geq c) = 1 - F_{\overline{D}}(c). \end{aligned}$$

The ROC curve is defined as the set of points $\{(\text{FPF}(c), \text{TPF}(c)) : c \in \mathbb{R}\}$ and, as it is clear from this definition, it lies in the unit square. Letting $t = \text{FPF}(c)$, the ROC curve can be alternatively expressed as $\{(t, \text{ROC}(t)) : t \in [0, 1]\}$, with

$$\text{ROC}(t) = 1 - F_D\{F_{\overline{D}}^{-1}(1 - t)\}.$$

ROC curves measure how separated the test outcomes in the diseased and nondiseased populations are (see Figure S1 of the Supplementary Materials). When the test outcomes in the two populations completely overlap, the ROC curve is the diagonal line of the unit square, that is, $\text{FPF}(c) = \text{TPF}(c)$ for all c, thus indicating a noninformative test. Conversely, the more separated the distributions of the test outcomes are, the closer the ROC curve is to the point $(0, 1)$ in the unit square and the better the diagnostic accuracy. A curve that reaches the point $(0, 1)$ has $\text{FPF}(c) = 0$ and $\text{TPF}(c) = 1$, for some threshold c and, hence, corresponds to a test that perfectly determines the true disease status.

It is common to summarize the information of the ROC curve into a single summary index and, undeniably, the most popular one is the area under the ROC curve (AUC), given by

$$\text{AUC} = \int_0^1 \text{ROC}(t)\text{d}t.$$

For a useless test that classifies individuals as diseased or nondiseased no better than chance, $\text{AUC} = 0.5$, whereas for a perfect test, $\text{AUC} = 1$. In addition to its geometric definition, the AUC has also a probabilistic interpretation,[7(p78)]

$$\text{AUC} = \Pr(Y_D \geq Y_{\overline{D}}),$$

that is, the AUC is the probability that the test outcome for a randomly chosen diseased subject exceeds the one exhibited by a randomly selected nondiseased individual.

2.2 | Modeling framework for the covariate-specific ROC curve

Let $\mathbf{X}$ denote the covariate vector and, for ease of notation, we will be assuming that the covariate vectors $\mathbf{X}_{\overline{D}}$ and $\mathbf{X}_D$ are the same in both populations. However, this is not necessarily always the case as, for instance, disease stage, which is a disease-specific covariate, might be of interest. The key object of our modeling framework is the covariate-specific ROC curve, which for a given covariate value $\mathbf{x}$, is defined as

$$\text{ROC}(t \,|\, \mathbf{x}) = 1 - F_D\{F_{\overline{D}}^{-1}(1 - t \,|\, \mathbf{x}) \,|\, \mathbf{x}\}, \quad 0 \leq t \leq 1, \tag{1}$$

where $F_D(y \,|\, \mathbf{x}) = \Pr(Y_D \leq y \,|\, \mathbf{X}_D = \mathbf{x})$ is the conditional cumulative distribution function in the diseased population, with $F_{\overline{D}}(y \,|\, \mathbf{x})$ being analogously defined. The covariate-specific counterpart of the AUC is given by

$$\text{AUC}(\mathbf{x}) = \int_0^1 \text{ROC}(t \,|\, \mathbf{x})\text{d}t. \tag{2}$$

Note that in this setting, for each possible value $\mathbf{x}$, we might obtain a different ROC curve/AUC and, therefore, also a possible different accuracy.

We follow an induced approach and we further assume that the relationship between covariates and test outcomes in each population is given by a location-scale regression model, that is,

$$Y_D = \mu_D(\mathbf{x}) + \sigma_D \varepsilon_D, \quad Y_{\overline{D}} = \mu_{\overline{D}}(\mathbf{x}) + \sigma_{\overline{D}} \varepsilon_{\overline{D}}, \tag{3}$$

where $\mu_D(\mathbf{x}) = E(Y_D \,|\, \mathbf{X}_D = \mathbf{x})$ and σ_D are the conditional mean function and scale parameter, respectively, in the diseased population; $\mu_{\overline{D}}(\mathbf{x})$ and $\sigma_{\overline{D}}$ are similarly defined. The errors ε_D and $\varepsilon_{\overline{D}}$ are independent of each other and independent of the covariates $\mathbf{X}_D$ and $\mathbf{X}_{\overline{D}}$, and in order to allow identifying the mean function and scale parameter in (3), are further assumed to have mean zero and unit variance. The corresponding cumulative distribution functions are denoted by F_{ε_D} and $F_{\varepsilon_{\overline{D}}}$, respectively. The independence between the error and the covariates in the location-scale regression model, allows one to rewrite the conditional cumulative distribution function of the test outcomes in the diseased population in terms of the cumulative distribution function of the regression errors in the same population, that is,

$$F_D(y \,|\, \mathbf{x}) = F_{\varepsilon_D}\left(\frac{y - \mu_D(\mathbf{x})}{\sigma_D}\right). \tag{4}$$

An analogous relationship holds between the conditional quantile function and the quantile function of the error terms, and namely, in the nondiseased population we have

$$F^{-1}_{\overline{D}}(1 - t \,|\, \mathbf{x}) = \mu_{\overline{D}}(\mathbf{x}) + \sigma_{\overline{D}} F^{-1}_{\varepsilon_{\overline{D}}}(1 - t). \tag{5}$$

Plugging in (4) and (5) into (1), the covariate-specific ROC curve can therefore be expressed as

$$\mathrm{ROC}(t \,|\, \mathbf{x}) = 1 - F_{\varepsilon_D}\left\{\frac{\mu_{\overline{D}}(\mathbf{x}) - \mu_D(\mathbf{x})}{\sigma_D} + \frac{\sigma_{\overline{D}}}{\sigma_D} F^{-1}_{\varepsilon_{\overline{D}}}(1 - t)\right\}, \quad 0 \le t \le 1.$$

An advantage of this formulation is that the cumulative distribution and quantile functions of the regression errors are not conditional, thus alleviating the computational burden. Note that under our formulation the effect of covariates on the ROC curve is expressed in terms of their effects on the mean functions of each population.

2.3 | Proposed robust and flexible estimator and its implementation

Let $\{(\mathbf{x}_{\overline{D}i}, y_{\overline{D}i})\}_{i=1}^{n_{\overline{D}}}$ and $\{(\mathbf{x}_{Dj}, y_{Dj})\}_{j=1}^{n_D}$ be two independent random samples of covariates and test outcomes from the nondiseased and diseased populations of size $n_{\overline{D}}$ and n_D, respectively. Further, for all $i = 1, \ldots, n_{\overline{D}}$ and $j = 1, \ldots, n_D$, let $\mathbf{x}_{\overline{D}i} = (x_{\overline{D}i,1}, \ldots, x_{\overline{D}i,p})'$ and $\mathbf{x}_{Dj} = (x_{Dj,1}, \ldots, x_{Dj,p})'$ be p-dimensional vectors of covariates.

2.3.1 | Modeling the mean function

From the location-scale regression models in (3), what needs to be specified is the regression function in each population. We will describe our modeling approach for the diseased population, but everything follows similarly for the nondiseased population. Since nonlinear relationships between test outcomes and continuous covariates often occur, we assume a flexible additive formulation for the mean function, namely,

$$\mu_D(\mathbf{x}_{Dj}) = \beta_{D0} + f_{D1}(x_{Dj,1}) + \ldots + f_{Dp}(x_{Dj,p}), \quad j = 1, \ldots, n_D,$$

where $f_{Dh}(\cdot)$, $h = 1, \ldots, p$, are smooth functions, each approximated by a linear combination of cubic B-splines basis functions defined over a sequence of knots $\xi_{Dh0} < \xi_{Dh1} < \ldots < \xi_{DhK_{Dh}} < \xi_{Dh,K_{Dh}+1}$.The knots ξ_{Dh0} and $\xi_{Dh,K_{Dh}+1}$ are boundary knots, while the remaining ones are interior knots. We then write

$$f_{Dh}(x_{Dj,h}) = \sum_{k=1}^{K_{Dh}+3} B_{Dhk}(x_{Dj,h})\beta_{Dhk} = \mathbf{B}'_{D\xi_{Dh}}(x_{Dj,h})\boldsymbol{\beta}_{Dh}, \quad j = 1, \ldots, n_D, \quad h = 1, \ldots, p,$$

where $\mathbf{B}_{D,\xi_{Dh}}(x_{Dj,h}) = (B_{Dh1}(x_{Dj,h}), \ldots, B_{Dh,K_{Dh}+3}(x_{Dj,h}))'$ with $B_{Dhk}(x)$ denoting the kth cubic B-spline basis function in the diseased population, evaluated at x, and defined by the knots sequence $\boldsymbol{\xi}_{Dh} = (\xi_{Dh0}, \xi_{Dh1}, \ldots, \xi_{Dh,K_{Dh}+1})'$, and $\boldsymbol{\beta}_{Dh} = (\beta_{Dh1}, \ldots, \beta_{Dh,K_{Dh}+3})'$. The mean function is thus expressed as

$$\begin{aligned}\mu_D(\mathbf{x}_{Dj}) &= \beta_{D0} + \mathbf{B}'_{D\xi_{D1}}(x_{Dj,1})\boldsymbol{\beta}_{D1} + \ldots + \mathbf{B}'_{D\xi_{Dp}}(x_{Dj,p})\boldsymbol{\beta}_{Dp} \\ &= \mathbf{z}'_{Dj}\boldsymbol{\beta}_D,\end{aligned} \tag{6}$$

where $\mathbf{z}'_{Dj} = (1, \mathbf{B}'_{D\xi_{D1}}(x_{Dj,1}), \ldots, \mathbf{B}'_{D\xi_{Dp}}(x_{Dj,p}))$ and $\boldsymbol{\beta}_D = (\beta_{D0}, \boldsymbol{\beta}'_{D1}, \ldots, \boldsymbol{\beta}'_{Dp})'$. It is well known that both the number and location of knots characterizing the B-splines basis functions are key choices that have the potential to impact the inferences, more so the former than the latter. As noted in Durrleman and Simon,[8] usually, only a few number of knots, say a maximum of three or four, are needed to adequately describe most of the phenomena likely to be observed in medical statistics. In this article, the selection of the number of knots is assisted by a robust version of the Akaike information criterion (see Section 2.4). Regarding the location of the K_{Dh} interior knots, we follow Rosenberg[9] and ξ_{Dhk} is set equal to the $k/(K_{Dh}+1)$ quantile of $\mathbf{x}_{D,h} = (x_{D1,h}, \ldots, x_{Dn_D,h})$, for $k = 1, \ldots, K_{Dh}$ and $h = 1, \ldots, p$, thus assuring an approximate equal number of observations at each interval defined by the knots. The boundary knots ξ_{Dh0} and $\xi_{Dh,K_{Dh}+1}$ are set equal to the minimum and maximum of $\mathbf{x}_{D,h}$, respectively. For the ease of presentation, we have assumed that all p covariates are continuous, but our modeling framework can also easily deal with categorical covariates, as well as, interactions between categorical covariates and interactions between a (smooth) continuous covariate and a categorical one.

2.3.2 | Robust estimation

The representation in (6) reduces the estimation of $\mu_D(\mathbf{x}_{Dj})$ to the estimation of the coefficient vector $\boldsymbol{\beta}_D$. Moreover, this expression is linear in $\boldsymbol{\beta}_D$, therefore allowing the use of well-established estimation techniques for multiple linear regression models. Estimation by ordinary least squares would be the most natural option. However, least squares type of approaches, because they rely on (minimizing) a quadratic loss function, are extremely sensitive to vertical outliers. Even a single atypical test outcome can drastically affect the estimated regression coefficients. Additionally, the scale parameter σ_D is traditionally estimated by the square root of $\widehat{\sigma}^2_D = (n_D - Q_D)^{-1}\sum_{j=1}^{n_D}(y_{Dj} - \mathbf{z}'_{Dj}\widehat{\boldsymbol{\beta}}^{\text{OLS}}_D)^2$, which is not robust either. Note that here Q_D is the dimension of the vector $\mathbf{z}_{Dj}$ and $\widehat{\boldsymbol{\beta}}^{\text{OLS}}_D$ is the least squares estimate of $\boldsymbol{\beta}_D$. It could be tempting to remove the outlying test outcomes using, for instance, graphical or residual analysis, and then obtaining the least squares estimates of the regression coefficients based on the "clean" sample. However, this strategy, might be not only impractical, but might also lead to inferences that are neither valid nor robust,[10] not to mention the reduction in sample size. One way to circumvent this problem is to minimize a less rapidly increasing function than the squared one, so that the influence of test outcomes with large residuals is reduced. For instance, least absolute deviation regression, which minimizes the absolute value loss function, $\sum_{j=1}^{n_D}|y_{Dj} - \mathbf{z}'_{Dj}\boldsymbol{\beta}_D|$, leads to estimators that are highly resistant to outliers (in the response variable). However, the drawback is that such estimators are relatively inefficient.[11(pp12,13)] An elegant compromise between the squared and absolute value loss functions was proposed by Huber,[12,13] who suggested to estimate $\boldsymbol{\beta}_D$ as

$$\widehat{\boldsymbol{\beta}}_D = \underset{\boldsymbol{\beta}_D}{\arg\min}\sum_{j=1}^{n_D}\rho\left(\frac{y_{Dj} - \mathbf{z}'_{Dj}\boldsymbol{\beta}_D}{\widehat{\sigma}_D}\right), \quad \rho(u) = \begin{cases} \frac{u^2}{2}, & |u| \le b_D, \\ |u|\ b_D - \frac{b_D^2}{2}, & |u| > b_D, \end{cases} \tag{7}$$

where $\widehat{\sigma}_D$ is a robust estimate of scale. The tuning constant b_D describes where the transition from a quadratic to a linear loss function takes place. Huber's loss function is quadratic for standardized residuals whose absolute value is equal or less than b_D and grows linearly for standardized residuals whose absolute values exceeds b_D. The parameter b_D therefore controls the amount of robustness. For larger values of b_D, Huber's loss function becomes more similar to the least squares loss function, whereas for small values of b_D, it is more similar to the absolute value loss function. Although b_D could be estimated from the data (as, eg, in Wang et al[14]), the typical choice of b_D is 1.345, for which the resulting estimator is asymptotically 95% as efficient as the least squares estimator when the true distribution of the errors is normal. In (7), the robust estimate of the scale $\widehat{\sigma}_D$, needed to ensure that the resulting estimate of $\boldsymbol{\beta}_D$ is scale equivariant, is in our case the rescaled median absolute deviation

$$\widehat{\sigma}_D = \underset{j=1,\ldots,n_D}{\text{median}}|y_{Dj} - \mathbf{z}'_{Dj}\widehat{\boldsymbol{\beta}}_D|/0.6745, \tag{8}$$

where the constant 0.6745, which corresponds to the 75th quantile of the standard normal distribution, is used to ensure that $\widehat{\sigma}_D$ is a consistent estimator of σ_D for normally distributed errors. Huber's estimator falls under the general

category of M-estimators (eg, Maronna et al[15(chaps2-5)]). The M-estimator minimizes (7) or, equivalently, solves the system of estimating equations

$$\sum_{j=1}^{n_D} \psi\left(\frac{y_{Dj} - \mathbf{z}'_{Dj}\boldsymbol{\beta}_D}{\widehat{\sigma}_D}\right)\mathbf{z}_{Dj} = \mathbf{0}_{Q_D}, \quad \psi(u) = \frac{\mathrm{d}}{\mathrm{d}u}\rho(u) = \begin{cases} u, & |u| \le b_D, \\ \text{sign}(u)\, b_D, & |u| > b_D, \end{cases} \tag{9}$$

where $\text{sign}(u) = I(u > 0) - I(u < 0)$, with $\text{sign}(0) = 0$, and $\mathbf{0}_{Q_D}$ denotes a vector of zeros of length Q_D. Defining the weight function $\omega(u)$ by

$$\omega(u) = \frac{\psi(u)}{u} = \begin{cases} 1, & |u| \le b_D, \\ \frac{b_D}{|u|}, & |u| > b_D, \end{cases}$$

allows us to rewrite Equation (9) as

$$\sum_{j=1}^{n_D} \omega_{Dj}\left(y_{Dj} - \mathbf{z}'_{Dj}\boldsymbol{\beta}_D\right)\mathbf{z}_{Dj} = \mathbf{0}_{Q_D}, \quad \omega_{Dj} = \omega\left(\frac{y_{Dj} - \mathbf{z}'_{Dj}\boldsymbol{\beta}_D}{\widehat{\sigma}_D}\right). \tag{10}$$

In Figure S2 of the Supplementary Materials, we present a comparison between Huber's ρ, ψ, and ω functions and the corresponding least squares and least absolute deviation counterparts for a better understanding of their behavior. Note that, for instance, least squares assigns equal weight to all observations, whereas Huber's based weight function assigns decreasing weights for observations with large, in absolute value, standardized residuals. The system of equations in (10) can be written in matrix form as

$$\mathbf{Z}'_D\boldsymbol{\Omega}_D\mathbf{Z}_D\boldsymbol{\beta}_D = \mathbf{Z}'_D\boldsymbol{\Omega}_D\mathbf{y}_D,$$

where $\mathbf{Z}_D$ is a matrix with $\mathbf{z}'_{Dj}$ as its jth row, $\boldsymbol{\Omega}_D$ is a diagonal matrix with entries given by ω_{Dj}, for $j = 1, \ldots, n_D$, and $\mathbf{y}_D = (y_{D1}, \ldots, y_{Dn_D})'$, and therefore can be regarded as a weighted least squares problem whose solution is given by $\widehat{\boldsymbol{\beta}}_D = (\mathbf{Z}'_D\boldsymbol{\Omega}_D\mathbf{Z}_D)^{-1}\mathbf{Z}'_D\boldsymbol{\Omega}_D\mathbf{y}_D$. Because the weights depend upon the estimated regression coefficients and scale parameter and, in turn, these depend upon the weights, the iteratively reweighted least squares procedure is employed. The algorithm can be briefly summarized by the following two steps.

Step 1: Obtain an initial estimate $\widehat{\boldsymbol{\beta}}_D^{(0)}$, which can be based, for instance, on a least squares fit. Use $\widehat{\boldsymbol{\beta}}_D^{(0)}$ to obtain $\widehat{\sigma}_D^{(0)}$ using the rescaled median absolute deviation as in (8). Compute an initial estimate of $\boldsymbol{\Omega}^{(0)}$ using $\widehat{\boldsymbol{\beta}}_D^{(0)}$ and $\widehat{\sigma}_D^{(0)}$.

Step 2: At iteration $k = 1, 2, \ldots$, solve for the new weighted least squares estimate $\widehat{\boldsymbol{\beta}}_D^{(k)} = (\mathbf{Z}'_D\boldsymbol{\Omega}_D^{(k-1)}\mathbf{Z}_D)^{-1}\mathbf{Z}'_D\boldsymbol{\Omega}_D^{(k-1)}\mathbf{y}_D$. This estimate will be used to obtain $\widehat{\sigma}_D^{(k)}$ and to compute $\boldsymbol{\Omega}_D^{(k)}$ which, in turn, will form the basis of $\widehat{\boldsymbol{\beta}}_D^{(k+1)}$. The iterative procedure is run until some convergence criterion is met.

The converged estimate $\widehat{\boldsymbol{\beta}}_D$ is taken as our final robust estimate of $\boldsymbol{\beta}_D$ and used to obtain the final estimate $\widehat{\sigma}_D$ of σ_D. We note here that $\widehat{\boldsymbol{\beta}}_D$ based on Huber's loss function is not robust against outliers in the covariates.

Once estimates $\widehat{\boldsymbol{\beta}}_D$ and $\widehat{\sigma}_D$ have been obtained, the distribution function of the error ε_D is estimated via a weighted empirical distribution function of the standardized residuals,

$$\widehat{F}_{\varepsilon_D}(y) = \frac{1}{\sum_{l=1}^{n_D}\omega^*_{Dl}}\sum_{j=1}^{n_D}\omega^*_{Dj} I\left(\widehat{\varepsilon}_{Dj} \le y\right), \quad \widehat{\varepsilon}_{Dj} = \frac{y_{Dj} - \widehat{\mu}_D(\mathbf{x}_{Dj})}{\widehat{\sigma}_D}, \quad \widehat{\mu}_D(\mathbf{x}_{Dj}) = \mathbf{z}^T_{Dj}\widehat{\boldsymbol{\beta}}_D, \quad \omega^*_{Dj} = \begin{cases} 1, & |\widehat{\varepsilon}_{Dj}| \le v_D, \\ 0, & |\widehat{\varepsilon}_{Dj}| > v_D. \end{cases} \tag{11}$$

The purpose of using a weighted version of the empirical distribution function of the standardized residuals is to downweight the influence of outlying test outcomes. Even if the regression coefficients and scale parameter are robustly estimated, the standardizing residuals corresponding to outlying observations may still lie far away from the bulk of test outcomes, therefore badly affecting the vanilla empirical estimate of the distribution function and consequently the estimate of the covariate-specific ROC curve and corresponding AUC. The tuning constant v_D in the weighting

function in (11)[11(p17)] controls whether an observation is retained or disregarded, that is, test outcomes whose standardized residuals, in absolute value, exceed v_D are completely eliminated in the weighting step. Using the normal distribution as a benchmark, $v_D = 3$ is deemed as reasonable.

Finally, the ROC curve estimate can be written as

$$\widehat{\mathrm{ROC}}(t\,|\,\mathbf{x}) = 1 - \widehat{F}_{\varepsilon_D}\left\{\frac{\widehat{\mu}_{\overline{D}}(\mathbf{x}) - \widehat{\mu}_D(\mathbf{x})}{\widehat{\sigma}_D} + \frac{\widehat{\sigma}_{\overline{D}}}{\widehat{\sigma}_D}\widehat{F}^{-1}_{\varepsilon_{\overline{D}}}(1-t)\right\}, \tag{12}$$

and the corresponding AUC admits the following closed-form expression, derived in the Appendix, and which can be regarded as a weighted robust covariate-specific Mann-Whitney type of statistic

$$\widehat{\mathrm{AUC}}(\mathbf{x}) = \frac{1}{\sum_{l=1}^{n_D}\omega^*_{Dl}\sum_{l=1}^{n_{\overline{D}}}\omega^*_{\overline{D}l}}\sum_{j=1}^{n_D}\sum_{i=1}^{n_{\overline{D}}}\omega^*_{Dj}\omega^*_{\overline{D}i}I\{\widehat{\mu}_{\overline{D}}(\mathbf{x}) + \widehat{\sigma}_{\overline{D}}\widehat{\varepsilon}_{\overline{D}i} \le \widehat{\mu}_D(\mathbf{x}) + \widehat{\sigma}_D\widehat{\varepsilon}_{Dj}\}. \tag{13}$$

Before proceeding it is worth mentioning that although all tuning constants were set using the standard normal distribution as a benchmark, the location-scale regression model in (3) only requires the error term to have zero mean and unit variance. However, if the error distribution is asymmetric, the choice of $b_D = 1.345$ induces bias at the intercept estimate and, consequently, the corresponding prediction of the conditional mean is also biased.[16,17] Increasing the tuning constant b_D reduces the bias but this parameter cannot be increased too much in order to maintain the robustness. At the time of writing of this article, we found the article by Fu and Wang[18] that tackles this problem by considering an asymmetric Huber loss function and which depends on two tuning constants that can be selected using the data-driven approach of Wang et al.[14] Nevertheless, from our computational experiences, the value of $b_D = 1.345$ is still somewhat reasonable (ie, it only leads to a very small amount of bias) for moderately asymmetric error distributions. With respect to the tuning parameter v_D, the value of, say 3, also works reasonably well for moderately asymmetric distributions. In practice, as a rule of thumb, we recommend to look at the histograms of the standardized residuals to check if the main bulk of these lie in the interval $[-3, 3]$. If not, the value of v_D should be changed accordingly. Also from our computational experiences, in most situations, adjusting the value of v_D suffices to obtain an unbiased estimate of the covariate-specific AUC, our main object of interest, even if the underlying estimate of the regression function is slightly biased (due to the bias in the estimate of the intercept). This shall be investigated in Section 3.

2.3.3 | Implementation

Some final comments on implementation are in order. The flexibility of our robust additive regression B-splines approach is controlled by selecting a small number of interior knots (say, a maximum of three or four), which we do with the aid of a robust version of the Akaike information criterion, as explained in the next section. Our procedure is easily implemented in `R`[19] using the `bs` function from the package `splines` (to create the cubic B-splines basis expansions) in combination with the `rlm` routine from the `MASS` package,[20] which performs the robust estimation procedure described above to obtain $\widehat{\boldsymbol{\beta}}_D$ and $\widehat{\sigma}_D$. The `R` code implementing our approach is publicly available at (https://github.com/vandainacio/robustROC). Of course, an alternative route would be M-estimation for additive models based on some form of regularization, where one starts off with many interior knots and uses a penalization based on some characteristic of the basis functions in order to control the smoothness of the fit (as, eg, in Wong et al[21]). However, such a procedure may involve intricate and computationally expensive algorithms. Because of this reason and also because in our experience a small number of knots usually suffices to describe the relationship between test outcomes and covariates, we regard our formulation based on robust additive regression B-splines to be a good balance between analytic and computational simplicity and the flexibility it affords.

2.4 | Robust Akaike information criterion

The issue of selecting the number of interior knots for each smooth function of a continuous covariate can be regarded as a model selection problem. Here, and because the classical Akaike information criterion (AIC) is sensitive to

outlying observations, such choice is assisted through the use of a robust version of the AIC, denoted by rAIC, that is suited for M-estimation and which was proposed by Tharmaratnam and Claeskens.[22] Specifically, the authors suggest to use

$$\text{rAIC}_D = 2\ n_D \log \widehat{\sigma}_D + 4\ \text{trace}(J_{D,n_D}^{-1} U_{D,n_D}), \tag{14}$$

where the empirical information matrices in the trace term (the penalty term) are calculated as follows

$$J_{D,n_D} = \frac{1}{n_D}\sum_{j=1}^{n_D} \psi'\left(\frac{y_{Dj} - \mathbf{z}'_{Dj}\widehat{\boldsymbol{\beta}}_D}{\widehat{\sigma}_D}\right)\frac{\mathbf{z}_{Dj}\mathbf{z}'_{Dj}}{\widehat{\sigma}_D^2}, \quad U_{D,n_D} = \frac{1}{n_D}\sum_{j=1}^{n_D} \psi^2\left(\frac{y_{Dj} - \mathbf{z}'_{Dj}\widehat{\boldsymbol{\beta}}_D}{\widehat{\sigma}_D}\right)\frac{\mathbf{z}_{Dj}\mathbf{z}'_{Dj}}{\widehat{\sigma}_D^2}.$$

Models with a varying number of interior knots will be fitted, $\boldsymbol{\beta}_D$ and σ_D are re-estimated in each model and the corresponding rAIC is computed, and the model with the smallest rAIC will be selected. When several continuous covariates are involved, our strategy involves exploring the set of all possible models. This is viable because not only is our fitting procedure extremely fast, but also because in medical diagnostic studies the number of continuous covariates available is often reduced and, as mentioned before, usually a modest number of knots suffices to describe the relationship between covariates and test outcomes. On a related task, the rAIC can also be used to select between a linear or a smooth effect of a given (continuous) covariate. It is important to remark that the penalty term needs to be changed to 2 trace$(J_{D,n_D}^{-1} U_{D,n_D})$ if instead of using the ψ function in (9), one uses 2 $\psi(u)$ (as, eg, in Tharmaratnam and Claeskens[22]).

2.5 | Bootstrap-based inference for the robust and flexible covariate-specific ROC curve

Confidence intervals for the covariate-specific ROC curve and corresponding AUC can be obtained through the bootstrap. We use a bootstrap of the residuals to resample the (robust) regression model in each population. The details of our bootstrap scheme are as follows. For $b = 1, \ldots, B$:

Step 1: Sample with replacement from the estimated standardized residuals $\{\widehat{\varepsilon}_{\overline{D}i}\}_{i=1}^{n_{\overline{D}}}$ and $\{\widehat{\varepsilon}_{Dj}\}_{j=1}^{n_D}$ to form bootstrap sets $\{\widehat{\varepsilon}^{(b)}_{\overline{D}i}\}_{i=1}^{n_{\overline{D}}}$ and $\{\widehat{\varepsilon}^{(b)}_{Dj}\}_{j=1}^{n_D}$.

Step 2: Use the mean function and variance estimates from the observed data to construct bootstrap samples $\{(\mathbf{x}_{\overline{D}i}, y^{(b)}_{\overline{D}i})\}_{i=1}^{n_{\overline{D}}}$ and $\{(\mathbf{x}_{Dj}, y^{(b)}_{Dj})\}_{j=1}^{n_D}$, where

$$y^{(b)}_{\overline{D}i} = \widehat{\mu}_{\overline{D}}(\mathbf{x}_{\overline{D}i}) + \widehat{\sigma}_{\overline{D}}\widehat{\varepsilon}^{(b)}_{\overline{D}i}, \quad y^{(b)}_{Dj} = \widehat{\mu}_D(\mathbf{x}_{Dj}) + \widehat{\sigma}_D\widehat{\varepsilon}^{(b)}_{Dj}.$$

Step 3: Repeat the estimation process with the bth bootstrap sample, thus obtaining $\widehat{\text{ROC}}^{(b)}(t\,|\,\mathbf{x})$ and $\widehat{\text{AUC}}^{(b)}(\mathbf{x})$.

Once this process has been completed, and according to the percentile method, a bootstrap confidence interval for, for example, AUC$(\mathbf{x})$, of confidence level $1-\alpha$ is given by

$$\left(\widehat{\text{AUC}}^{\alpha/2}(\mathbf{x}), \widehat{\text{AUC}}^{1-\alpha/2}(\mathbf{x})\right),$$

where $\widehat{\text{AUC}}^{\tau}(\mathbf{x})$ represents the τth percentile of the ensemble of estimates $\{\widehat{\text{AUC}}^{(b)}(\mathbf{x})\}_{b=1}^{B}$. We acknowledge that by "naively" resampling the standardized residuals, some bootstrap samples may have a proportion of outliers higher than the contamination level tolerated by our procedure. Nonetheless, some computational experiments (not shown) have demonstrated that even for the case of a contamination level of 15% in each population, the resulting 95% bootstrap confidence intervals for the covariate-specific AUC still have a coverage probability reasonably close to the nominal value. We note in passing that in diagnostic studies it is unlikely, in our experience, to have contamination percentages higher than 5% and for this reason we regard this bootstrap scheme to be viable in practice. Further, this bootstrap has been justified by Shorack,[23] with the only difference being that this author's approach does not bootstrap the scale parameter.

3 | SIMULATION STUDY

To evaluate the empirical performance of our robust and flexible approach for conducting inference about the covariate-specific AUC, we analyzed simulated data under four different scenarios (described in the next section). For each scenario, 1000 data sets were generated using sample sizes of $(n_{\overline{D}}, n_D) = (100,100)$, $(n_{\overline{D}}, n_D) = (200,100)$, and $(n_{\overline{D}}, n_D) = (200,200)$. The following percentages of test outcomes contamination, in each population, were considered: 2%, 5%, and 10%. The case of no contamination (original simulated datasets) was also considered in order to ascertain the performance of our method when a robust approach is not needed at all. Also, the case of a contamination of 10% is included mainly as a proof of concept because, as we have already mentioned, in our experience it is unlikely to encounter such a high contamination percentage in practice.

3.1 | Simulation scenarios

In Scenario I, we consider different homoscedastic linear mean regression models for the nondiseased and diseased populations, namely,

$$y_{\overline{D}i} = 0.5 + x_{\overline{D}i,1} + 1.5\varepsilon_{\overline{D}i}, \quad y_{Dj} = 2 + 4x_{Dj,1} + 2\varepsilon_{Dj}, \quad i = 1, \ldots, n_{\overline{D}}, \quad j = 1, \ldots, n_D.$$

The primary purpose of including this scenario is to allow us assessing the impact of using a cubic B-splines basis formulation for the mean function of each population when the underlying true effect is, in fact, linear. Data for Scenario II are governed by the following nonlinear mean regression models

$$y_{\overline{D}i} = \sin\{\pi x_{\overline{D}i,1}\} + 0.5\varepsilon_{\overline{D}i}, \quad y_{Dj} = 1 + x_{Dj,1}^2 + \varepsilon_{Dj}.$$

Scenario III involves heteroscedastic nonlinear mean regression models for the diseased and nondiseased populations

$$y_{\overline{D}i} = \sin\{\pi x_{\overline{D}i,1}\} + (1 + 0.75x_{\overline{D}i,1})\varepsilon_{\overline{D}i}, \quad y_{Dj} = 1 + x_{Dj,1}^2 + (1 + x_{Dj,1})\varepsilon_{Dj}.$$

Note that our model is actually misspecified in this case as it does not allow the variance to change with the covariates and the goal of including this scenario is exactly to assess the performance of our approach when the assumption of constant variance does not hold. Finally, in Scenario IV, we have considered the case where two continuous covariates affect the test outcomes

$$y_{\overline{D}i} = 0.5 + x_{\overline{D}i,1} + x_{\overline{D}i,2}^2 + 1.5\varepsilon_{\overline{D}i}, \quad y_{Dj} = 2 + 4x_{Dj,1}^3 + 1.5x_{Dj,2} + 2\varepsilon_{Dj}.$$

In all cases, the continuous covariates x_1 and x_2, are independently generated from uniform distributions, namely,

$$x_{\overline{D}i,1} \overset{\text{i.i.d.}}{\sim} \text{U}(0,1), \quad x_{\overline{D}i,2} \overset{\text{i.i.d.}}{\sim} \text{U}(0,2), \quad x_{Dj,1} \overset{\text{i.i.d.}}{\sim} \text{U}(0,1), \quad x_{Dj,2} \overset{\text{i.i.d.}}{\sim} \text{U}(0,2),$$

and, in addition,

$$\varepsilon_{\overline{D}i} \overset{\text{i.i.d.}}{\sim} \text{N}(0,1), \quad \varepsilon_{Dj} \overset{\text{i.i.d.}}{\sim} \text{N}(0,1).$$

Further, in all scenarios, the contaminated data were generated by randomly selecting a given percentage of test outcomes and replacing them by $y_{\overline{D}} = \mu_{\overline{D}}(\mathbf{x}_{\overline{D}}) + \kappa_{\overline{D}}\sigma_{\overline{D}}(\mathbf{x}_{\overline{D}}) + \sigma_{\overline{D}}(\mathbf{x}_{\overline{D}})\varepsilon_{\overline{D}}$ and $y_D = \mu_D(\mathbf{x}_D) + \kappa_D\sigma_D(\mathbf{x}_D) + \sigma_D(\mathbf{x}_D)\varepsilon_D$ (shift in the location outliers) in the nondiseased and diseased populations, respectively, and where the covariates and error terms follow the same distributions as the noncontaminated data. Note that for all scenarios but the third we have $\sigma_{\overline{D}}(\mathbf{x}_{\overline{D}}) \equiv \sigma_{\overline{D}}$ and $\sigma_D(\mathbf{x}_D) \equiv \sigma_D$. Additionally, we have considered $\kappa_{\overline{D}} = 15$ and $\kappa_D = 20$, which at a first glance might seem excessive but it is indeed in line with what we observe in our data application in Section 4 (see also the left panel of Figure 3). The impact of the magnitude of those values on the estimates will be discussed in Section 3.3.

3.2 | Models

For each simulated dataset, we fit our approach considering no interior knots for each continuous covariate in each population (ie, $K_{\overline{D}1} = K_{\overline{D}2} = K_{D1} = K_{D2} = 0$). A further inspection to this choice is discussed in the next section. Our model is compared to the semiparametric approach of Pepe,[3] which is based on a location-scale regression model for the test outcomes in each population that relies on a linear formulation for the mean function and with the regression coefficients estimated, for instance, by least squares. In addition to the original approach proposed by Pepe,[3] we have also considered an extension of this method by using a cubic B-splines trend, also with no interior knots, so that direct comparisons to our approach are easier and fairer. The only difference between ours and this approach is the objective function (least squares vs Huber's ρ function). In addition, our method is also compared to the nonparametric approach of Rodríguez-Álvarez et al,[5] which relies on kernel-based estimators for the mean and variance functions of the location-scale model. The main difference of this approach to the one of González-Manteiga et al[4] is the order of the local polynomial smoothers used for estimating the regression function; while González-Manteiga et al[4] employed a local constant fit (order 0), Rodríguez-Álvarez et al[5] considered a linear fit (order 1). Because local constant regression suffers from boundary-bias problems, we only considered the latter approach. All competing methods were implemented using the `ROCnReg` package[24] which, in turn, relies on the `np` package[25] for kernel estimation. Still on the kernel method, it is important to note that the bandwidth parameters involved in the estimation process were selected using least-squares cross-validation. We further remark that the kernel approach, as it stands now, can only deal with one continuous covariate.

3.3 | Results

The case $(n_{\overline{D}}, n_D) = (200,100)$, which is similar to the prostate cancer application in Section 4, is shown here and we first analyze Scenarios I to III. The estimated (mean across the 1000 Monte Carlo estimates) covariate-specific AUC along with the 2.5% and 97.5% simulation quantiles in Figure 1 illustrate the ability of our model to accurately and precisely capture complex functional forms in a case where the contamination in each population is 5%. As can be observed in Figure 1, the three non-robust estimators have a very poor performance, showing some bias and wide simulation quantiles bands. Further, and obviously, the original estimator proposed by Pepe[3] is inadequate for scenarios involving nonlinear trends. Also, note that in Scenario III, where the underlying regression models in the two populations are heteroscedastic, our estimator still has a very decent performance, although we expect it to deteriorate for more substantial changes in the variance along with the covariate. We further note that the kernel approach is the only one tailored for such a heterocedastic scenario.

The remaining sample sizes and percentages of contamination are shown in Figures S3 to S14 in the Supplementary Materials and although similar conclusions were found, some comments are in order. First, even in the case of no contamination, Figures S3, S7, and S11 in the Supplementary Materials, corresponding, respectively, to Scenarios I, II, and III, the performance of our estimator is basically on par with that of the non-robust and flexible estimators. Second, in the case of a 2% contamination (see Figures S4, S8, and S12 in the Supplementary Materials), the non-robust estimators already show some bias and an increase in the width of the simulation bands. This is, of course, much more marked for the case of 10% contamination. In turn, the performance of our robust estimator remains quite good.

For Scenario IV, which involves two continuous covariates, only our estimator was considered. We regard this scenario mainly as a proof of concept when there are multiple continuous covariates and the results obtained from fitting the competing approaches were similar to those reported for Scenarios I to III. Nonetheless, for the three sample sizes and different percentages of contamination considered, our approach performs very well and is able to recover the different profiles of the true covariate-specific surface (Figure 2 and Figures S15-S18 in the Supplementary Materials).

We shall remark that although the covariate-specific AUC admits the closed-form expression in (13), its calculation can be very time-consuming, especially for large datasets. As a consequence, here and in Section 4, the integral in (2) was approximated using Simpson's rule. In our experience, Simpson's rule provides almost identical results to the ones obtained using the closed-form expression.

Because we rely on the robust AIC to assist in the selection of the number of knots needed to appropriately model the regression function, we have investigated the behavior of this criterion when performing such a task. Specifically, over the

1000 simulated datasets, for each scenario considered, for the different sample sizes in each population (100 and 200) and for the different contamination percentages, we computed the percentage over the 1000 simulation runs that the robust AIC favored the model with no interior knots over a model with three interior knots. For this latter model, following the rule discussed in Section 2, the knots are located at the 0.25, 0.5, and 0.75 quantiles of the covariates. Note that for Scenario IV, as a slight simplification, we have assumed the same number of knots for both continuous covariates (ie, $(K_{D1}, K_{D2}) = (0, 0)$ and $(K_{D1}, K_{D2}) = (3, 3)$, with the same applying in the nondiseased population). Results are displayed in Tables 1 to 4 in the Supplementary Materials and show that, most of the time, the robust AIC favored the simpler model with no interior knots over the more complex model with three interior knots. For instance, in Scenario I, where the regression function assumes a linear form in both populations, our intuition would dictate that the model with no interior knots should be selected for a large number of the simulated datasets and Table S1 (Supplementary Materials) confirms exactly this. Also, in Scenario 4, the model with no interior knots for the two covariates (and that involves seven regression parameters) is favored most of the time over the model that uses three interior knots for each of the covariates (and that involves thirteen regression parameters).

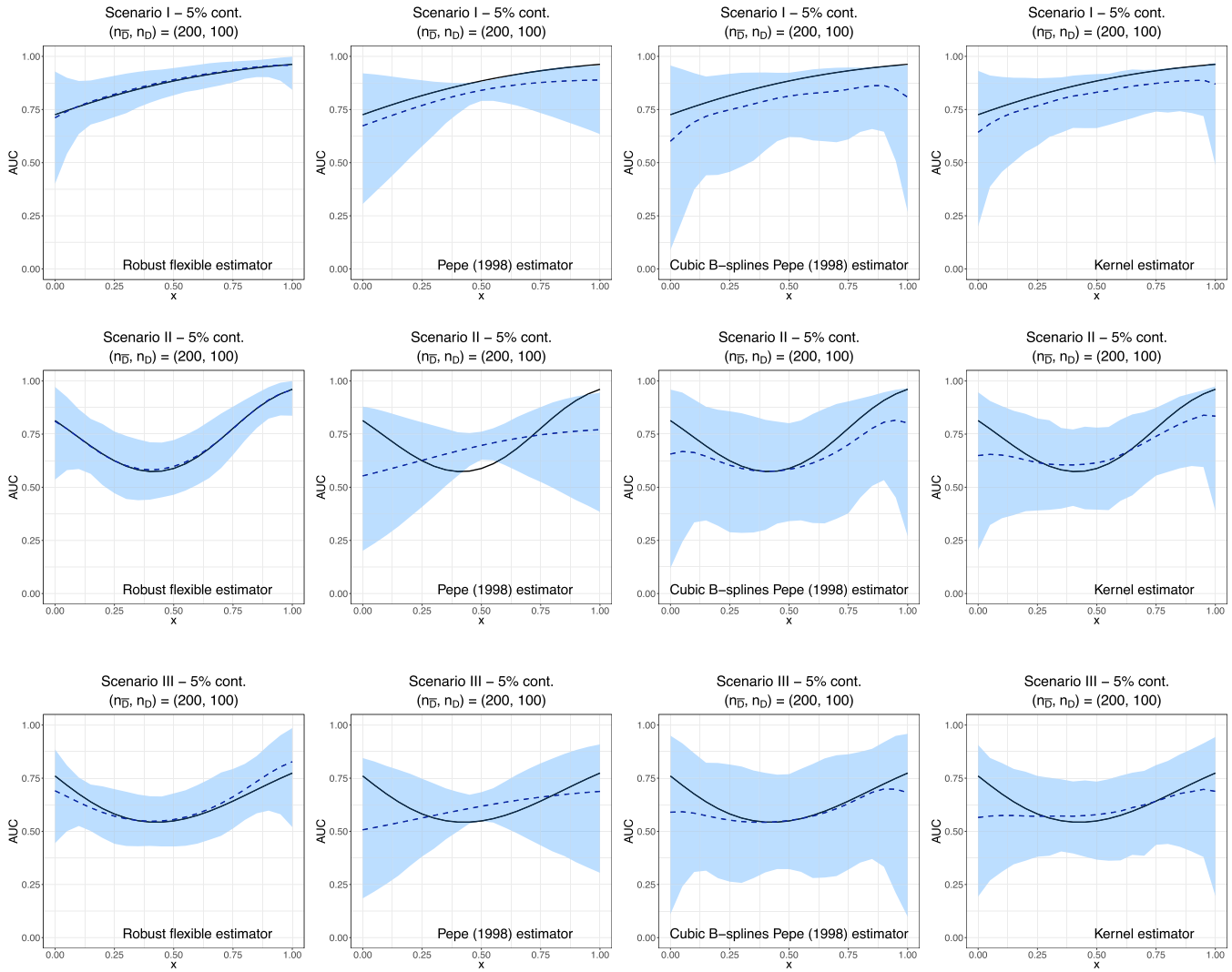

FIGURE 1 True covariate-specific AUC (solid line) vs the mean of the Monte Carlo estimates (dashed line) along with the 2.5% and 97.5% simulation quantiles (shaded area) for the case of 5% contamination. The first row displays the results for Scenario I, the second row for Scenario II, and the third row for Scenario III. The first column corresponds to our flexible and robust estimator, the second column to the estimator proposed by Pepe,[3] the third one to the cubic B-splines extension of Pepe,[3] and the fourth column to the kernel estimator. For all scenarios $(n_{\overline{D}}, n_D) = (200,100)$ [Colour figure can be viewed at wileyonlinelibrary.com]

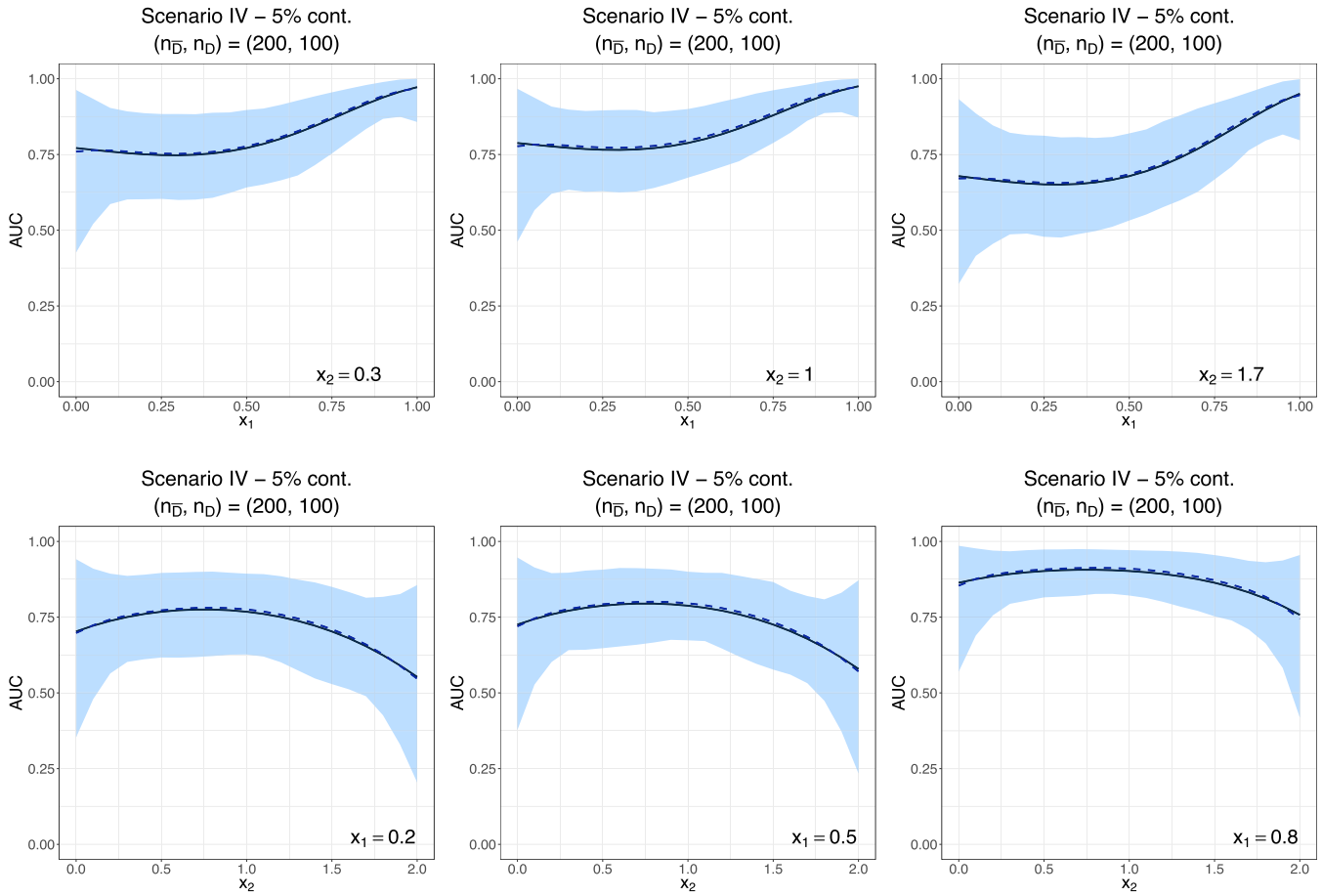

FIGURE 2 Scenario IV. Multiple profiles of the true covariate-specific AUC (solid line) vs the mean of the Monte Carlo estimates (dashed line) along with the 2.5% and 97.5% simulation quantiles (shaded area) for the case of 5% contamination and for $(n_{\overline{D}}, n_D) = (200,100)$ [Colour figure can be viewed at wileyonlinelibrary.com]

We conclude this section with some extra important remarks. Although we have assumed that both populations were subject to contamination, it may happen that only test outcomes from one of the populations are contaminated. Simulation results (not shown) indicate that in such cases the robust estimator still outperforms the non-robust competitors. However, and interestingly, even when assuming balanced sample sizes, contamination in the nondiseased population seems to impact much more the ability of the non-robust estimators to recover the true functional form of the AUC than contamination in the diseased population. Our intuitive explanation, bearing in mind Equation (12), is that estimation of the quantile function of the standardized residuals is more impacted by outliers than the estimation of the cumulative distribution function (of the standardized residuals). Further, a shift of $15\sigma_{\overline{D}}(\mathbf{x}_{\overline{D}})$ and $20\sigma_D(\mathbf{x}_D)$ in the location of the distribution of the test outcomes in the nondiseased and diseased populations, respectively, was considered. Our computational experiments (results not shown) revealed that the performance of the non-robust estimators is affected by the magnitude of those shifts and, as expected, the larger the shift, the worse the performance. On the other hand, the performance of our robust estimator is basically unchanged. We have, however, noticed that if the outliers are too small, in magnitude (eg, by considering a shift smaller than 5 times the standard deviation), they might pass unnoticed when computing the weighted empirical distribution function of the standardized residuals (see (11)), and this causes some bias for contaminations close to 10% and onwards. Apart from the case just mentioned where outliers result from very small shifts in the mean of the distribution, our computational experiments also revealed that with contamination percentages of about 15% and onwards (in each population), the performance of our estimator starts deteriorating. Finally, we should also mention that having also simulated contaminated samples considering radial outliers, which arise by multiplying the scale of the distribution of test outcomes in each group by a given factor, results remained basically the same and therefore are not shown here.

In the Supplementary Materials, we replicate Scenarios I and II for the case where the error term in each population follows a two-component (symmetric) mixture of normal distributions and also the case where it follows a skewed distribution. In the latter case, and for a highly skewed error distribution, slightly adjusting the value of v_d, $d \in \{D, \overline{D}\}$ was enough to obtain reasonable estimates of the covariate-specific AUC. We have also considered a simulation scenario where the covariate-specific AUC shows a marked nonlinearity, with the goal of checking whether a few number of interior knots suffice to accurately recover the true form such conditional AUC. The results shown in Figures S25 and S26 of the Supplementary Materials show that this was indeed the case.

4 | APPLICATION

4.1 | Motivation and exploratory analysis

Prostate cancer (PCa) is the second most frequent cancer diagnosed in men, only after lung cancer, and amounts to the fifth highest cause of death worldwide.[26] Gleason histological scoring system is the most reliable system used for the grading of prostate cancer, but it requires invasive tissue biopsies. This, and the rising incidence of prostate cancer worldwide, have led to the search of less invasive biomarkers that can accurately predict the presence of PCa. The PHI, that combines three prostate specific antigen subforms into a single score using a mathematical formula, has been introduced[27] and since then several studies have shown that it significantly improves prediction of a positive biopsy when compared to the prostate specific antigen.[28-30] The PHI is now approved by the US Food and Drug Administration and it has also been adopted into the US National Cancer Network guidelines. We apply our methods to data from a study designed to assess the added value of the PHI to multi-parametric magnetic resonance imaging in detecting significant prostate cancers (Gleason $\geq$ 7) in a repeat biopsy population.[31] Here our goal is slightly distinct and we seek to assess, if and how, the ability of the PHI to discriminate between men with benign or Gleason 6 PCa (which throughout we refer as the nondiseased group and for which $n_{\overline{D}} = 185$) and men with Gleason 7 or above PCa (which we term as the diseased group and for which $n_D = 94$), changes with age. To the best of our knowledge, this is the first attempt to study the possible age effect on the accuracy of the PHI to distinguish between those two PCa groups. In Figure 3 (left panel), we show the histograms of the PHI levels in the two populations and it can be observed that, as expected, men belonging to the group defined by Gleason $\geq$ 7 tend to have higher PHI values than those with a benign lesion or with a Gleason of 6. We can also notice that although the majority of PHI values lie below 100 in the nondiseased group and below 150 in the diseased group, there are two PHI scores, one from each group, above 200.

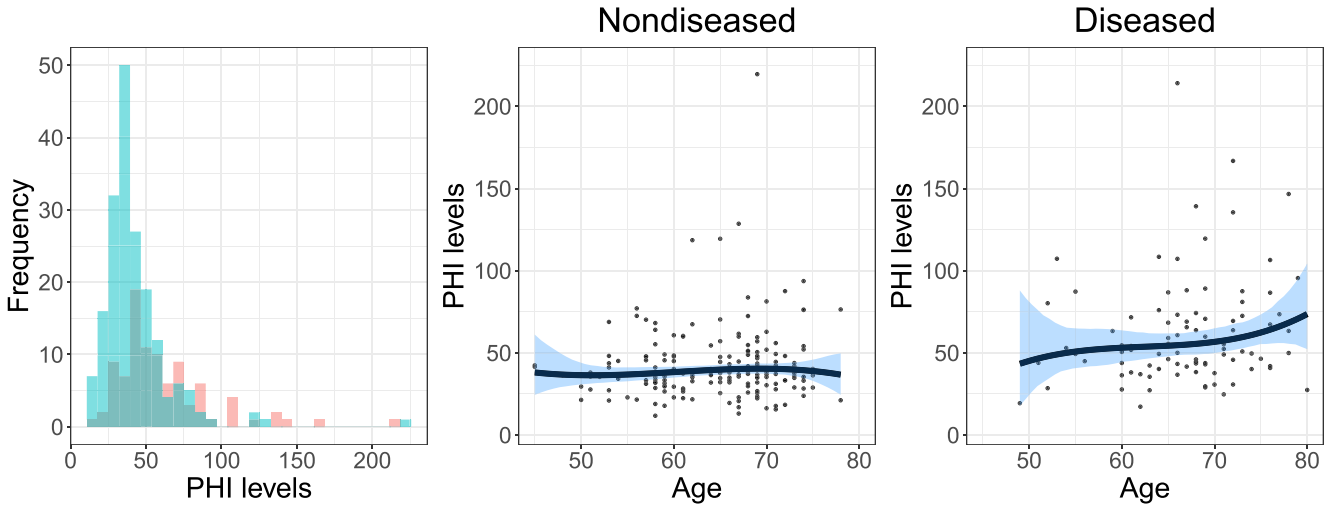

FIGURE 3 Left panel: Histogram of the PHI scores from the nondiseased (blue) and diseased (red) populations. Middle and right panels: Regression functions resulting from fitting our approach. The solid line is the point estimate, while the shaded areas represent the 95% pointwise bootstrap confidence bands (based on 1000 resamples) [Colour figure can be viewed at wileyonlinelibrary.com]

4.2 | Unconditional and age-specific ROC analysis

We start our analysis by calculating the AUC when ignoring the potential age effect and we have computed it in a robust way (so that it is more easily comparable to the covariate-specific AUCs we will present later in this section) as

$$\widehat{\text{AUC}} = \frac{1}{\sum_{l=1}^{n_D}\omega^*_{Dl}\sum_{l=1}^{n_{\overline{D}}}\omega^*_{\overline{D}l}}\sum_{j=1}^{n_D}\sum_{i=1}^{n_{\overline{D}}}\omega^*_{Dj}\omega^*_{\overline{D}i}\left\{I(y_{\overline{D}i} < y_{Dj}) + \frac{1}{2}I(y_{\overline{D}i} = y_{Dj})\right\},$$

where the weights $\omega^*_{\overline{D}i}$ and ω^*_{Dj} are defined similarly as in (11) and arise from fitting, in each group, a robust regression model with the PHI scores as the responses and with only an intercept term. Although PHI outcomes are defined on a continuous scale, in practice ties can occur, and so the extra term $(1/2) \times I(y_{\overline{D}i} = y_{Dj})$ corrects for such possible ties. The resulting AUC estimate (95% bootstrap confidence interval based on 1000 resamples) is 0.74 (0.67, 0.82), revealing a reasonably good capacity of the PHI levels to discriminate between men with a Gleason of 6 or a benign lesion and men with Gleason ≥ 7.

We now turn our attention to the inclusion of age in the analysis. In Figure 3, middle and right panels are depicted the scatter plots of the data in each group along with the estimated regression functions; the robust AIC in (14) led to $K_{D1} = K_{\overline{D}1} = 0$ (no interior knots), with these selected from the set $\{0, 1, 2, 3, 4\}$. First, both scatter plots do not indicate any departure from the homoscedasticity assumption. Second, as a result of the weighting scheme in (10) behind the estimation of the regression coefficients, such high PHI values do not push the regression functions toward them as much as the analogous least squares counterparts (shown in Figure S27 of the Supplementary Materials). Note that for a fairer comparison we have also included, in Figure S27 of the Supplementary Materials, an approach that models the mean function through a cubic B-splines basis expansion with no interior knots. Third, while in the nondiseased group the PHI does not show any noticeable dynamic along age, in the diseased group there seems to be slight evidence that older ages are associated with higher PHI outcomes. In Figure 4 (left and middle panels), we present two different age-specific ROC curves, namely, for ages of 57 and 73 years old, with the corresponding AUCs being 0.71 (0.53, 0.88) and 0.79 (0.67, 0.90), respectively. As can be seen, the ROC curves are somewhat jagged, which is due to the fact of them being based on the (weighted) empirical distribution function of the standardized residuals. To inspect the age effect further, Figure 4 (right panel) shows a plot of the age-specific AUC for ages between 55 and 75 years old and we can observe that the capacity of the PHI levels to distinguish between men with benign or Gleason 6 PCa and men with Gleason ≥ 7 PCa slightly increases with age, ranging from 0.70 (0.50, 0.89) for a men of 55 years old to 0.84 (0.71, 0.94) for a men of 75 years old. The AUC estimate obtained when ignoring the age effect was 0.74 and so, roughly, for individuals younger than 70 years we would be slightly overestimating the accuracy of the PHI scores and for individuals older than 70 years old such accuracy would

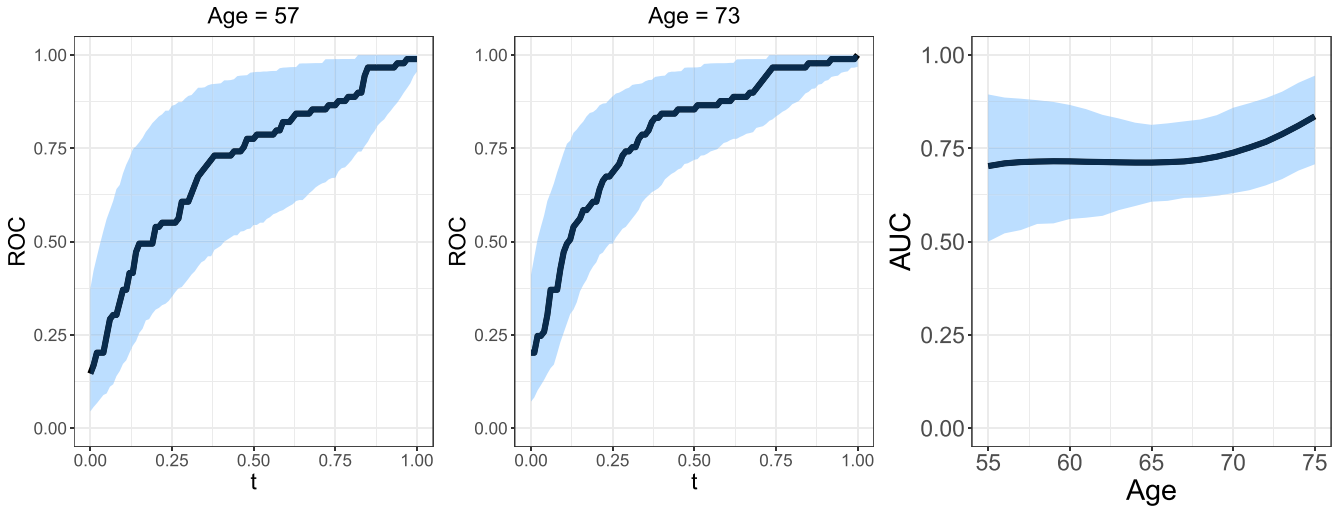

FIGURE 4 Left and middle panels: Two age-specific ROC curves. Right panel: Age-specific AUC. The solid line is the point estimate, while the shaded areas represent the 95% pointwise bootstrap confidence bands (based on 1000 resamples) [Colour figure can be viewed at wileyonlinelibrary.com]

be slightly underestimated. Nonetheless, note that the unconditional AUC estimate and corresponding 95% confidence interval are contained in the 95% bootstrap confidence band for all ages considered and so it is difficult to draw firm conclusions about the age effect. We remark that AUC predictions were only considered for ages in the interval (55, 75) as this corresponds to the range where both groups had a reasonable number of observations. We further remark that when computing the 95% bootstrap confidence bands, the number of internal knots selected for the observed data (in this case this was 0 for both groups) was used when recomputing the estimates for the generated bootstrap samples. Also we highlight that it took less than one minute to run our model (including the 1000 bootstrap resamples) on a MacBook Pro with 2.3 GHz Intel i5 processor and 8 GB RAM. At this point it is fair to remark that both the unconditional and conditional results were obtained under the choice of $\nu_{\overline{D}} = \nu_D = 3$. By looking at the histogram of the standardized residuals in each of the populations (Figure S28 in the Supplementary Materials), such a choice seems to be reasonable. Leaving apart the clear outlying test outcomes (those above 200), only less than 5% of the outcomes in each population were above 3 or below −3 and so considering $\nu_{\overline{D}} = \nu_D = 4$ and $\nu_{\overline{D}} = \nu_D = 5$ made no difference (Figure S28 of the Supplementary Materials). Finally, in Figure S29 of the Supplementary Materials, we present the age-specific AUC estimates obtained when considering the three non-robust estimators detailed in Section 3, and as can be observed they are not markedly different from the point estimate provided by our approach. This should come as no surprise as the estimated mean functions were also not too distinct, which makes sense as there are only two PHI outcomes, one in each group, that lie well above the remaining scores. Also, all approaches agree that the accuracy of the PHI scores to distinguish between the two groups of PCa slightly increases with age.

5 | CONCLUDING REMARKS

We have developed a robust and flexible modeling framework for estimating the covariate-specific ROC curve and corresponding AUC that assumes a location-scale regression model in both the diseased and nondiseased populations and that combines an additive regression B-splines formulation with M-estimation for the mean function. Additionally, a weighted version of the empirical distribution function of the standardized residuals is used to estimate the distribution function of the error term. Our approach is thus able to simultaneously accommodate outlying test outcomes and nonlinear effects of the covariates. The proposed methodology has the additional appealing features of being simple and computationally inexpensive. The simulation study conducted illustrated the ability of our method to recover the true shape of the covariate-specific ROC curve and AUC in a variety of complex scenarios involving different test outcome distributions and contamination percentages. Simulation results also show that although our approach works best under symmetric error distributions, it can still deal decently with moderately skewed error distributions. Our investigation into the potential of the PHI to distinguish between men with a benign lesion or a Gleason 6 prostate cancer and men with aggressive prostate cancer (Gleason 7 or above) found that its accuracy slightly increases with age. Although in this particular case the overall message of our analysis agrees with that provided by the non-robust estimators, our approach enabled us to identify one outlying test outcome in each population.

Once a diagnostic test/biomarker has proved to have a desired discriminatory ability, the next step is to determine which cutoff value to use to diagnose/screen subjects in practice and this threshold value may depend on covariates as well. Our method can be trivially adapted to also estimate the covariate-specific Youden index and its corresponding optimal threshold. In particular, since

$$\mathrm{YI}(\mathbf{x}) = \max_c \{F_{\overline{D}}(c\,|\,\mathbf{x}) - F_D(c\,|\,\mathbf{x})\}, \tag{15}$$

one can make use of the result in (4) and estimate the cumulative distribution function of the standardized residuals using (11). The covariate-specific optimal threshold is the one maximizing (15). However, when the cumulative distribution functions in (15) are those based on the empirical distribution function of the standardized residuals (see (4)), the resulting covariate-specific threshold curves have the drawback of being too jagged, especially for small sample sizes, which may be unappealing for practitioners.

Finally, throughout we have assumed that only the test outcomes were prone to outliers. However, if covariates are also contaminated, our approach can be easily extended to cope with this case by considering MM-estimation techniques instead of the M-estimation method used here.

ACKNOWLEDGEMENTS

Vanda Inácio and Miguel de Carvalho: Partially funded by Fundação para a Ciência e Tecnologia (Portugal) through projects: PTDC/MAT-STA/28649/2017 and UID/MAT/00006/2020. Vanda Lourenço: Partially funded by Fundação para a Ciência e Tecnologia (Portugal) through projects: UIDB/00297/2020 and SFRH/BSAB/142919/2018. VML also acknowledges Erasmus Mobility Grants 29191/002/2017/STT, 29191/036/2018/STT and 032/2020/SAM/FCT.

DATA AVAILABILITY STATEMENT

Upon publication, the R code will be available from an online publicly repository together with a sample simulated data. The real data due to confidentiality reasons cannot be shared.

ORCID

Vanda Inácio https://orcid.org/0000-0001-8084-1616
Vanda M. Lourenço https://orcid.org/0000-0001-8338-7279
Miguel de Carvalho https://orcid.org/0000-0003-3248-6984

REFERENCES

1. Inácio V, Rodríguez-Álvarez MX, Gayoso-Diz P. Statistical evaluation of medical tests. *Ann Rev Stat Appl*. 2021;8:41-67.
2. Racine JS. *Reproducible Econometrics Using R*. Oxford, UK: Oxford University Press; 2019.
3. Pepe MS. Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics*. 1998;54(1):124-135.
4. González-Manteiga W, Pardo-Fernández JC, Keilegom IV. ROC curves in non-parametric location-scale regression models. *Scand J Stat*. 2011;38(1):169-184.
5. Rodríguez-Álvarez MX, Roca-Pardiñas J, Cadarso-Suárez C. ROC curve and covariates: extending induced methodology to the non-parametric framework. *Stat Comput*. 2011;21(4):483-499.
6. Rodríguez A, Martínez JC. Bayesian semiparametric estimation of covariate-dependent ROC curves. *Biostatistics*. 2014;15(2):353-369.
7. Pepe MS. *The Statistical Evaluation of Medical Tests for Classification and Prediction*. Oxford, UK: Oxford University Press; 2003.
8. Durrleman S, Simon R. Flexible regression models with cubic splines. *Stat Med*. 1989;8(5):551-561.
9. Rosenberg PS. Hazard function estimation using B-splines. *Biometrics*. 1995;51(3):874-887.
10. Welsh AH, Ronchetti E. A journey in single steps: robust one-step M-estimation in linear regression. *J Stat Plan Infer*. 2002;103(1-2):287-310.
11. Rousseeuw PJ, Leroy AM. *Robust Regression and Outlier Detection*. Hoboken, NJ: John Wiley & Sons; 1987.
12. Huber PJ. Robust estimation of a location parameter. *Ann Math Stat*. 1964;45(1):73-101.
13. Huber PJ. Robust regression: asymptotics, conjectures and Monte Carlo. *Ann Stat*. 1973;1(5):799-821.
14. Wang YG, Lin X, Zhu M, Bai Z. Robust estimation using the Huber function with a data-dependent tuning constant. *J Comput Graph Stat*. 2007;16(2):468-481.
15. Maronna RA, Martin RD, Yohai VJ, Salibián-Barrera M. *Robust Statistics: Theory and Methods (with R)*. Hoboken, NJ: John Wiley & Sons; 2019.
16. Fan J, Li Q, Wang Y. Estimation of high dimensional mean regression in the absence of symmetry and light tail assumptions. *J Royal Stat Soc Ser B*. 2017;79(1):247.
17. Wang L, Zheng C, Zhou W, Zhou WX. A new principle for tuning-free Huber regression. *Stat Sin*. 2020.
18. Fu L, Wang YG. Robust regression with asymmetric loss functions. *Stat Methods Med Res*. 2021.
19. R Core Team. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing; 2020.
20. Venables WN, Ripley BD. *Modern Applied Statistics with S*. 4th ed. New York, NY: Springer; 2002.
21. Wong RK, Yao F, Lee TC. Robust estimation for generalized additive models. *J Comput Graph Stat*. 2014;23(1):270-289.
22. Tharmaratnam K, Claeskens G. A comparison of robust versions of the AIC based on M-S-and MM-estimators. *Statistics*. 2013;47(1):216-235.
23. Shorack GR. Bootstrapping robust regression. *Commun Stat Theory Methods*. 1982;11(9):961-972.
24. Rodríguez-Álvarez MX, Inacio V. ROCnReg: An R Package for Receiver Operating Characteristic Curve Inference with and without Covariates, The R Journal, 2021;13(1):525-555.
25. Hayfield T, Racine JS. Nonparametric econometrics: the np package. *J Stat Softw*. 2008;27(5):1-32.
26. Rawla P. Epidemiology of prostate cancer. *World J Oncol*. 2019;10(2):63.
27. Le BV, Griffin CR, Loeb S, et al. [-2] Proenzyme prostate specific antigen is more accurate than total and free prostate specific antigen in differentiating prostate cancer from benign disease in a prospective prostate cancer screening study. *J Urol*. 2010;183(4):1355-1359.
28. Stephan C, Vincendeau S, Houlgatte A, Cammann H, Jung K, Semjonow A. Multicenter evaluation of [- 2] proprostate-specific antigen and the prostate health index for detecting prostate cancer. *Clin Chem*. 2013;59(1):306-314.

29. Wang W, Wang M, Wang L, Adams TS, Tian Y, Xu J. Diagnostic ability of % p2PSA and prostate health index for aggressive prostate cancer: a meta-analysis. *Sci Rep*. 2014;4:5012.
30. De La Calle C, Patil D, Wei JT, et al. Multicenter evaluation of the prostate health index to detect aggressive prostate cancer in biopsy naive men. *J Urol*. 2015;194(1):65-72.
31. Gnanapragasam V, Burling K, George A, et al. The prostate health index adds predictive value to multi-parametric MRI in detecting significant prostate cancers in a repeat biopsy population. *Sci Rep*. 2016;6(1):1-8.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Inácio V, M. Lourenço V, de Carvalho M, Parker RA, Gnanapragasam V. Robust and flexible inference for the covariate-specific receiver operating characteristic curve. *Statistics in Medicine*. 2021;40(26):5779-5795. https://doi.org/10.1002/sim.9153

APPENDIX A. WEIGHTED ROBUST COVARIATE-SPECIFIC AUC

Here we deduce the representation of our weighted robust covariate-specific AUC in the form of (13). The derivation is based on simple calculus and its main steps are outlined below. We start by noting that

$$\begin{aligned}\widehat{\mathrm{AUC}}(\mathbf{x}) &= \int_0^1 \widehat{\mathrm{ROC}}(t\,|\,\mathbf{x})\mathrm{d}t \\ &= \int_0^1 \left[1 - \widehat{F}_{\varepsilon_D}\left\{\frac{\widehat{\mu}_{\overline{D}}(\mathbf{x}) - \widehat{\mu}_D(\mathbf{x})}{\widehat{\sigma}_D} + \frac{\widehat{\sigma}_{\overline{D}}}{\widehat{\sigma}_D}\widehat{F}^{-1}_{\varepsilon_{\overline{D}}}(1-t)\right\}\right]\mathrm{d}t \\ &= \int_0^1 \sum_{j=1}^{n_D} \frac{\omega^*_{Dj}}{\sum_{l=1}^{n_D}\omega^*_{Dl}} I\left\{\widehat{\varepsilon}_{Dj} \geq \frac{\widehat{\mu}_{\overline{D}}(\mathbf{x}) - \widehat{\mu}_D(\mathbf{x})}{\widehat{\sigma}_D} + \frac{\widehat{\sigma}_{\overline{D}}}{\widehat{\sigma}_D}\widehat{F}^{-1}_{\varepsilon_{\overline{D}}}(1-t)\right\}\mathrm{d}t,\end{aligned}$$

which implies that

$$\begin{aligned}\widehat{\mathrm{AUC}}(\mathbf{x}) &= \frac{1}{\sum_{l=1}^{n_D}\omega^*_{Dl}}\sum_{j=1}^{n_D}\omega^*_{Dj}\int_0^1 I\left\{t \geq 1 - \widehat{F}_{\varepsilon_{\overline{D}}}\left(\frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_{\overline{D}}(\mathbf{x})}{\widehat{\sigma}_{\overline{D}}} + \frac{\widehat{\sigma}_D}{\widehat{\sigma}_{\overline{D}}}\widehat{\varepsilon}_{Dj}\right)\right\}\mathrm{d}t \\ &= \frac{1}{\sum_{l=1}^{n_D}\omega^*_{Dl}}\sum_{j=1}^{n_D}\omega^*_{Dj}\sum_{i=1}^{n_{\overline{D}}}\frac{\omega^*_{\overline{D}i}}{\sum_{l=1}^{n_{\overline{D}}}\omega^*_{\overline{D}l}} I\left\{\widehat{\varepsilon}_{\overline{D}i} \leq \frac{\widehat{\mu}_D(\mathbf{x}) - \widehat{\mu}_{\overline{D}}(\mathbf{x})}{\widehat{\sigma}_{\overline{D}}} + \frac{\widehat{\sigma}_D}{\widehat{\sigma}_{\overline{D}}}\widehat{\varepsilon}_{Dj}\right\} \\ &= \frac{1}{\sum_{l=1}^{n_D}\omega^*_{Dl}\sum_{l=1}^{n_{\overline{D}}}\omega^*_{\overline{D}l}}\sum_{j=1}^{n_D}\sum_{i=1}^{n_{\overline{D}}}\omega^*_{Dj}\omega^*_{\overline{D}i} I\{\widehat{\mu}_{\overline{D}}(\mathbf{x}) + \widehat{\sigma}_{\overline{D}}\widehat{\varepsilon}_{\overline{D}i} \leq \widehat{\mu}_D(\mathbf{x}) + \widehat{\sigma}_D\widehat{\varepsilon}_{Dj}\}.\end{aligned}$$