Biostatistics (2012), **13**, 4, *pp.* 594–608
doi:10.1093/biostatistics/kxs007
Advance Access publication on April 2, 2012

Extending induced ROC methodology to the functional context

VANDA INÁCIO*
Department of Statistics and Operations Research and Center of Statistics and Applications, Lisbon University, Lisbon 1749, Portugal
vanda.kinets@gmail.com

WENCESLAO GONZÁLEZ-MANTEIGA, MANUEL FEBRERO-BANDE
Department of Statistics and Operations Research, Faculty of Mathematics,University of Santiago de Compostela, Santiago de Compostela 15782, Spain

FRANCISCO GUDE
Clinical Epidemiology Unit, Hospital Clínico Universitario de Santiago and Instituto de Investigación Sanitária de Santiago (IDIS), Santiago de Compostela, Santiago de Compostela 15782, Spain

TODD A. ALONZO
Division of Biostatistics, University of Southern California, California, CA 91066, USA

CARMEN CADARSO-SUÁREZ
Department of Statistics and Operations Research, Faculty of Medicine and Instituto de Investigación Sanitária de Santiago (IDIS), Santiago de Compostela, Spain

Summary

The receiver operating characteristic (ROC) curve is the most widely used measure for evaluating the discriminatory performance of a continuous marker. Often, covariate information is also available and several regression methods have been proposed to incorporate covariate information in the ROC framework. Until now, these methods are only developed for the case where the covariate is univariate or multivariate. We extend ROC regression methodology for the case where the covariate is functional rather than univariate or multivariate. To this end, semiparametric- and nonparametric-induced ROC regression estimators are proposed. A simulation study is performed to assess the performance of the proposed estimators. The methods are applied to and motivated by a metabolic syndrome study in Galicia (NW Spain).

Keywords: Area under the curve; Functional data; Functional linear model; Functional nonparametric model; Metabolic syndrome; ROC curve.

1. Introduction

The receiver operating characteristic (ROC) curve is a popular method for evaluating the performance of continuous markers and its presence is widespread in medical studies. Parametric and nonparametric

*To whom correspondence should be addressed.

estimators are available (Zou *and others*, 1997; Pepe, 2003; Peng and Zhou, 2004, among others). Often, covariate information that affects the marker performance is also available and ignoring such covariates can yield biased or oversimplified inferences. Various methods have been proposed to assess possible covariate effects on the ROC curve. Induced methodology is based on using separate regression models for the healthy and diseased populations and then computing the induced form of the ROC curve (Pepe, 1998; Faraggi, 2003; Zheng and Heagerty, 2004; González-Manteiga *and others*, 2011; Rodríguez-Álvarez *and others*, 2011b). Alternatively, direct methodology assumes a regression model for the ROC curve itself, with the effects of the covariates being directly evaluated on the ROC curve (Alonzo and Pepe, 2002; Pepe, 2003; Cai, 2004). For a comparative study of both methodologies, see Rodríguez-Álvarez *and others* (2011a). Until now, these methodologies are only developed for the cases where the covariate is univariate or multivariate, although in some settings of practical interest, the covariate can have a more complex structure.

In this paper, we extend the estimation of the conditional ROC curve for the cases where the covariate is functional within the induced context. A functional covariate means that the explanatory variable is valued in an infinite-dimensional space. Examples of functional variables include minute by minute values of a speculative asset, meteorological and pollution monitoring data, seismic data, growth curves and heart rates, and a plethora of examples in all fields of science and engineering. Analyzing functional data with standard multivariate methods that ignore the functional nature of the data may significantly impact the inferences. Thus, there is a need for specific techniques that can handle such data and extract relevant information from it. For an overview of this topic see, for instance, Ramsay and Silverman (2006) or Ferraty and Vieu (2006). Our approach is motivated by a medical study, conducted in Galicia—Spain, concerning the use of the gamma-glutamyl-transferase (GGT) as a diagnostic test to detect women with metabolic syndrome. Recent investigations suggest an association between the GGT levels and nocturnal hypoxemia (decrease in arterial oxygen saturation). To this end, the arterial oxygen saturation was measured every 20 s during the patient's sleep. It is our aim to investigate how the discriminatory ability of GGT to detect metabolic syndrome is affected by the oxygen saturation. We should remark that our approach is different from the existing approach for longitudinal markers (Etzioni *and others*, 1999; Zheng and Heagerty, 2004). First, longitudinal and functional data are different in nature. While in the context of longitudinal data analysis, a random function typically represents a subject that is often observed at a small number of time points; in the functional setup, the data are recorded densely over time, often by a machine (Zhao *and others*, 2004). Second, in longitudinal studies, the concepts of sensitivity and specificity incorporate both the time-varying nature of the marker and the clinical onset time of the disease, whereas in our case, sensitivity and specificity are not time dependent and the subjects do not change the state of health during the study.

For the most appropriate analysis of a given set of data, one desires a variety of readily available models from which the data analyst can choose the most appropriate approach. In this work, we present two estimators of the induced covariate-specific ROC curve: (1) a semiparametric estimator based on a homocedastic functional linear model and (2) a nonparametric estimator based on an extension to the functional context of kernel regression techniques. The functional linear model has been popularized by Ramsay and Delzel (1991) and it imposes a linear constraint on the regression relationship. The linear constraint is useful when the curves are very distant from each other, but it may be inappropriate for some applications (Yao and Müller, 2010). It is therefore of interest to consider a flexible alternative. The functional nonparametric model (see Ferraty and Vieu, 2002 for a first study of this model and Ferraty and Vieu, 2006 for an overview) has been studied recently and is an interesting and complementary alternative to the functional linear model and is based solely on the assumption that the effect of the continuous covariate follows a smooth function.

The remainder of the paper is organized as follows. In Section 2, we provide background material on functional regression models. The estimation procedures used to derive the induced functional covariate-specific ROC curve are presented in Section 3. In Section 4, a simulation study is carried out to assess the performance of the proposed models. Application of the proposed methods to a real example concerning the diagnosis of metabolic syndrome is presented in Section 5. Concluding remarks are given in Section 6.

2. BACKGROUND

2.1 *Functional linear model with scalar response*

In the simplest setting, the functional predictor and the scalar response are related by a linear operator. Given a scalar response y on $\mathbb{R}$ and a smooth random predictor process X on a compact support T that is square integrable (i.e., $\int_T X^2(t)\mathrm{d}t < \infty$), the classical functional linear model relates y and X by

$$y = \langle X, \beta\rangle + \varepsilon = \int_T X(t)\beta(t)\mathrm{d}t + \varepsilon. \tag{2.1}$$

Here, $\langle\cdot,\cdot\rangle$ denotes the usual inner product on $L^2(T)$, the separable Hilbertian space of square integrable functions defined on T, the regression parameter function β is also assumed to be smooth and square integrable, and ε is a real random variable with zero mean and finite variance σ^2, and such that $\mathbb{E}[X(t)\varepsilon] = 0$ for $t \in T$. For simplicity, we assume that both variables are centered, i.e. $\mathbb{E}[X(t)] = 0$ for $t \in T$ and $\mathbb{E}[y] = 0$.

Suppose now we observe a random sample $\{(y_i, X_i)\}_{i=1}^n$. The model in (2.1) suggests to estimate β by minimizing the residual sum of squares

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n}(y_i - \langle X_i, \beta\rangle)^2, \tag{2.2}$$

which may be accomplished by the principal components approach developed by Cardot *and others* (1999) and further analyzed by Cai and Hall (2006) and Febrero-Bande *and others* (2010), among others. This estimation method works as follows. Let Γ_X be the sample covariance operator of $X_1, \ldots, X_n$ which transforms any function Z in $L^2(T)$ into another function in $L^2(T)$, given by

$$\Gamma_X Z = \frac{1}{n}\sum_{i=1}^{n}\langle X_i, Z\rangle X_i.$$

The sample covariance operator Γ_X admits a spectral decomposition in terms of the orthonormal eigenfunctions $\{v_k\}_{k=1,2,\ldots}$, which forms a complete basis of the functional space, with associated nonnegative and nondecreasing eigenvalues $\{\lambda_k\}_{k=1,2,\ldots}$, such that $\Gamma_X v_k = \lambda_k v_k$ for $k \geqslant 1$. By the well-known Karhunen–Loève expansion, the predictor process admits the following representation:

$$X_i(t) = \sum_{k=1}^{\infty}\gamma_{ik}v_k,$$

where $\gamma_{ik} = \langle X_i, v_k\rangle$ are the principal component scores. Recall that the regression parameter function β is square integrable and $\{v_k\}_{k=1,2,\ldots}$ form a complete orthonormal basis, we have

$$\beta(t) = \sum_{k=1}^{\infty}\beta_k v_k(t),$$

where $\beta_k = \langle \beta, \upsilon_k \rangle$ for $k \geqslant 1$. Using these expansions, we can write the residual sum of squares in (2.2) as

$$\mathrm{RSS}(\beta) = \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{\infty} \gamma_{ik} \beta_k \right)^2 .$$

As noted in Ramsay and Silverman (2006), minimizing such sum of squares, with respect to $\beta_1, \beta_2, \ldots$, yields a regression estimator that adapts perfectly to the sample points but is not very informative. To tackle this problem, Cardot *and others* (1999) proposed to estimate β by taking $\beta_k = 0$, for $k \geqslant k_n + 1$, where k_n is some positive integer such that $k_n < n$ and $\lambda_{k_n} > 0$, and estimating the coefficients β_k, for $k = 1, \ldots, k_n$, by minimizing the residual sum of squares given by

$$\mathrm{RSS}(\beta_{(1:k_n)}) = \sum_{i=1}^{n} \left(y_i - \sum_{k=1}^{k_n} \gamma_{ik} \beta_k \right)^2 = \| Y - \gamma_{(1:k_n)} \beta_{(1:k_n)} \|^2,$$

where $Y = (y_1, \ldots, y_n)'$, $\beta_{(1:k_n)}$ is the k_n-vector $\beta_{(1:k_n)} = (\beta_1, \ldots, \beta_{k_n})'$, and $\gamma_{(1:k_n)}$ is the $n \times k_n$ matrix whose kth column is the vector $\gamma_{.k} = (\gamma_{1k}, \ldots, \gamma_{nk})'$, the kth principal component score, which verifies $\gamma'_{.k}\gamma_{.k} = n\lambda_k$ and $\gamma'_{.k}\gamma_{.l} = 0$ for $k \neq l$. Using standard arguments, the least squares estimate of $\beta_{(1:k_n)}$ is

$$\widehat{\beta}_{(1:k_n)} = \left(\frac{\gamma'_{.1} Y}{n\lambda_1}, \ldots, \frac{\gamma'_{.k_n} Y}{n\lambda_{k_n}} \right),$$

which, finally, allows us to write the least squares estimate of the slope β, denoted by $\widehat{\beta}_{(k_n)}$

$$\widehat{\beta}_{(k_n)} = \sum_{k=1}^{k_n} \widehat{\beta}_k \upsilon_k = \sum_{k=1}^{k_n} \frac{\gamma'_{.k} Y}{n\lambda_k} \upsilon_k. \tag{2.3}$$

To determine the cutoff k_n in an automatic data-driven way, we have chosen the cutoff that minimizes the predictive cross validation (PCV) criterion

$$\mathrm{PCV}(k) = \sum_{i=1}^{n} (y_i - \langle X_i, \widehat{\beta}_{(-i,k)} \rangle)^2, \quad k = 1, \ldots, k_{\max}, \tag{2.4}$$

where $\widehat{\beta}_{(-i,k)}$ is the least squares estimate of β using the cutoff k and leaving out the ith observation (y_i, X_i) in the estimation. We pick the k that minimizes this criterion.

2.2 *Functional nonparametric model*

The functional nonparametric model is an interesting and complementary alternative to the functional linear model. Moreover, when dealing with functional data, it is difficult to gain an intuition on whether the linear model is adequate or which parametric model would best fit the data. We focus on the following functional nonparametric regression model:

$$Y = \mu(X) + \varepsilon,$$

where Y is a scalar response variable, X is a functional covariate, μ is an unknown but smooth regression function, and the error ε satisfies $\mathbb{E}[\varepsilon | X] = 0$ and $\mathbb{E}[\varepsilon^2 | X] = \sigma^2 < \infty$.

Nadaraya (1964) and Watson (1964) proposed to estimate μ as a locally weighted average using a kernel as a weighting function. Recently, Ferraty and Vieu (2006) extended to the functional context the Nadaraya–Watson estimator; they proposed to estimate μ as

$$\widehat{\mu} = \sum_{i=1}^{n} W_{i,h}(X) y_i, \quad \text{with } W_{i,h}(X) = \frac{K(h^{-1}d(X, X_i))}{\sum_{i=1}^{n} K(h^{-1}d(X, X_i))}, \tag{2.5}$$

where $X_1, \ldots, X_n$ are a sample of curves for which the corresponding responses $y_1, \ldots, y_n$ have been observed and X is an additional fixed curve. Additionally, K is an asymmetric decreasing kernel function, h is a positive smoothing parameter or bandwidth, and d is a suitable semimetric in the functional space.

It is easy to see that the weights $W_{i,h}(X)$ in (2.5) sum up to one and therefore the estimator is a weighted average of the y_is. It is clear that the smaller $d(X, X_i)$, the larger $K(h^{-1}d(X, X_i))$, i.e. the closer X_i is to X, the larger is the weight assigned to y_i. The parameter h plays a major role because it controls the amount of weighting given to the y_is. The smaller h is, the more $\widehat{\mu}(X)$ is sensitive to small variations of the y_is. In the opposite case, the larger h is, the larger is the weight assigned to distant observations. In other words, if h is too small, the estimator will be too rough; but if it is too large, important features will be smoothed out. In Fig. 1 of the supplementary material available at *Biostatistics* online, we show the effect of the bandwidth on the ROC curve estimate. Cross validation (CV) is a popular method to automatically select h. The criterion is

$$\text{CV}(h) = \sum_{i=1}^{n} (y_i - \widehat{\mu}_h^{-i}(X_i))^2, \tag{2.6}$$

where $\widehat{\mu}_h^{-i}(X_i)$ indicates the estimate at X_i leaving out the ith element of the sample. We pick the h that minimizes this criterion.

The choice of the semimetric is also crucial to the performance of the estimator and must be related to the particular features of the data set at hand. Ferraty and Vieu (2006, p 223) suggest to choose the semimetric based on the smoothness or roughness of the predictor curves $X_1, \ldots, X_n$. Specifically, when the curves are smooth, they suggest to use the L_2-norm of the qth derivative of the curve, while for rough curves, they recommend semimetrics based on principal component analysis. For the definition of this class of semimetrics, see Ferraty and Vieu (2006, p 28–30).

Regarding the choice of the kernel, any sensible choice will produce acceptable results, and thus this choice is much less important than the choice of h and the semimetric.

3. Induced ROC regression methodologies

3.1 *Regression model*

A location-scale regression model is assumed for the marker result in both healthy and diseased populations. More specifically, let

$$Y_{\bar{D}} = \mu_{\bar{D}}(X) + \sigma_{\bar{D}}(X)\varepsilon_{\bar{D}} \quad \text{and} \quad Y_D = \mu_D(X) + \sigma_D(X)\varepsilon_D,$$

where X denotes the functional covariate, $\mu_{\bar{D}}$ and μ_D are the regression functions, and $\sigma^2_{\bar{D}}$ and σ^2_D are the variance functions. The errors $\varepsilon_{\bar{D}}$ and ε_D are independent of each other with zero mean, unit variance, and distributions $G_{\bar{D}}$ and G_D, respectively. To guard against misspecification of error distributions, we do not assume specific distributions for the errors.

Based on the above location-scale regression models, the covariate-specific ROC curve can be expressed as

$$\mathrm{ROC}_X(p) = 1 - G_D\left(G_{\bar{D}}^{-1}(1-p)\frac{\sigma_{\bar{D}}(X)}{\sigma_D(X)} - \frac{\mu_D(X) - \mu_{\bar{D}}(X)}{\sigma_D(X)}\right), \quad 0 < p < 1$$

where $G_{\bar{D}}^{-1}(1-p) = \inf\{y\colon G_{\bar{D}}(y) \geqslant 1-p\}$. The most popular summary measure of the diagnostic accuracy is the area under the curve (AUC), which is given by

$$\mathrm{AUC}_X = \int_0^1 \mathrm{ROC}_X(p)\mathrm{d}p.$$

For an useless test, AUC $= 0.5$, and for a perfect test, AUC $= 1$.

3.2 *Semiparametric induced ROC regression estimator*

We extend to the functional context the semiparametric model of Pepe (1998). In this model, the variance parameters are not allowed to depend on covariates, i.e. we are dealing with homocedastic linear models. In such case, $\sigma_{\bar{D}}(X) = \sigma_{\bar{D}}$, $\sigma_D(X) = \sigma_D$, and the covariate effect on the ROC curve is contained in the covariate effect on the difference in means between diseased and nondiseased subjects, $\mu_{\bar{D}}(X) - \mu_D(X)$, where $\mu_{\bar{D}}(X) = \langle X, \beta_{\bar{D}}\rangle$ and $\mu_D = \langle X, \beta_D\rangle$.

Let $\{(y_{\bar{D}i}, X_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_{Dj}, X_{Dj})\}_{j=1}^{n_D}$ be two independent random samples of size $n_{\bar{D}}$ and n_D from the healthy and diseased populations, respectively. The estimation procedure is as follows:

1. Estimate $\beta_{\bar{D}}$ and β_D using the estimators proposed in (2.3) on the basis of samples $\{(y_{\bar{D}i}, X_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_{Dj}, X_{Dj})\}_{j=1}^{n_D}$, respectively; as in Section 2.1, we are assuming that both the covariate and the test result are centered.
2. Estimate $\sigma_{\bar{D}}^2$ and σ_D^2 as

$$\widehat{\sigma}_{\bar{D}}^2 = \frac{\sum_{i=1}^{n_{\bar{D}}}(y_{\bar{D}i} - \widehat{\mu}_{\bar{D}}(X_i))^2}{n_{\bar{D}} - k_{n_{\bar{D}}} - 1} \quad \text{and} \quad \widehat{\sigma}_D^2 = \frac{\sum_{j=1}^{n_D}(y_{Dj} - \widehat{\mu}_D(X_j))^2}{n_D - k_{n_D} - 1},$$

   where $k_{n_{\bar{D}}}$ and k_{n_D} are the number of principal components chosen by the PCV criterion.
3. Estimate distribution functions $G_{\bar{D}}$ and G_D on the basis of the empirical distribution of the standardized residuals

$$\widehat{G}_{\bar{D}}(y) = \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} I\left[\frac{y_{\bar{D}i} - \widehat{\mu}_{\bar{D}}(X_i)}{\widehat{\sigma}_{\bar{D}}} \leqslant y\right], \quad \widehat{G}_D(y) = \frac{1}{n_D}\sum_{j=1}^{n_D} I\left[\frac{y_{Dj} - \widehat{\mu}_D(X_j)}{\widehat{\sigma}_D} \leqslant y\right].$$

4. Calculate the covariate-specific ROC curve as follows:

$$\widehat{\mathrm{ROC}}_X(p) = 1 - \widehat{G}_D\left(\widehat{G}_{\bar{D}}^{-1}(1-p)\frac{\widehat{\sigma}_{\bar{D}}}{\widehat{\sigma}_D} - \frac{\widehat{\mu}_D(X) - \widehat{\mu}_{\bar{D}}(X)}{\widehat{\sigma}_D}\right), \quad 0 < p < 1.$$

3.3 *Nonparametric induced ROC regression estimator*

While the semiparametric approach has the advantage of being more efficient if the parametric form of the model is correct, it may misspecify the correct model form. Nonparametric models provide an alternative solution and are more robust and data adaptive. Within this context, the robustness is achieved by means of not assuming any parametric forms for the mean and variance functions, and the errors can depend heterocedastically on the functional covariate through $\sigma_{\bar{D}}$ and σ_D. The proposed estimation scheme is the following:

1. Estimate the regression functions $\mu_{\bar{D}}$ and μ_D using the kernel smoother estimator proposed in (2.5).
2. Estimate the variance functions $\sigma^2_{\bar{D}}$ and $\sigma^2_{\bar{D}}$, using the transformed samples $\{(z_{Di}, X_{Di})\}_{i=1}^{n_D}$ and $\{(z_{\bar{D}i}, X_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$, where $z_{Di} = (y_{Di} - \widehat{\mu}(X_{Di}))^2$ and $z_{\bar{D}i} = (y_{\bar{D}i} - \widehat{\mu}(X_{\bar{D}i}))^2$.
3. Estimate the distribution functions $G_{\bar{D}}$ and G_D by the empirical distribution of the standardized residuals

$$\widehat{G}_{\bar{D}}(y) = \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} I\left[\frac{y_{\bar{D}i} - \widehat{\mu}_{\bar{D}}(X_i)}{\widehat{\sigma}_{\bar{D}}(X_i)} \leqslant y\right], \quad \widehat{G}_D(y) = \frac{1}{n_D}\sum_{j=1}^{n_D} I\left[\frac{y_{Dj} - \widehat{\mu}_D(X_j)}{\widehat{\sigma}_D(X_j)} \leqslant y\right].$$

4. Finally, compute the covariate-specific ROC curve as follows:

$$\widehat{\mathrm{ROC}}_X(p) = 1 - \widehat{G}_D\left(\widehat{G}^{-1}_{\bar{D}}(1-p)\frac{\widehat{\sigma}_{\bar{D}}(X)}{\widehat{\sigma}_D(X)} - \frac{\widehat{\mu}_D - \widehat{\mu}_{\bar{D}}(X)}{\widehat{\sigma}_D(X)}\right), \quad 0 < p < 1$$

4. SIMULATION STUDY

In this section, we present the results of a simulation study conducted to evaluate the small sample performance of the proposed methods. Two different simulation scenarios were considered, namely: (a) a linear scenario and (b) a nonlinear scenario, which, from an applied standpoint, appears plausible given the results obtained in the data analysis (see Section 5).

Each predictor trajectory was observed discretely over the domain [0, 1] on an equally spaced grid of $N = 51$ points and all of them were generated with a trend function $X_0(t) = t + \sin(t), 0 < t < 1$, and a covariance function derived from two eigenfunctions $\upsilon_1(t) = \sqrt{2}\sin(0.5\pi t)$ and $\upsilon_2(t) = \sqrt{2}\sin(1.5\pi t)$ associated with eigenvalues $\lambda_1 = 4$, $\lambda_2 = 1$, as well as $\lambda_m = 0$, $m \geqslant 3$. The predictor functional principal component (FPC) scores are $\gamma_m \sim \mathrm{N}(0, \lambda_m)$, $m = 1, 2$. In short, the predictor trajectories were generated using the following:

$$X(t) = X_0(t) + \sum_{m=1}^{2} \gamma_m \upsilon_m(t). \tag{4.7}$$

Figure 2(a) in supplementary material available at *Biostatistics* online gives an idea of their shape. In both scenarios, the response was generated from a single regression function $\beta(t) = \upsilon_1(t) + \upsilon_2(t)$. The scenarios considered were as follows:

- *Scenario 1*

$$Y_D = 2 + 1.5\langle\beta, X_D\rangle + 2\varepsilon_D, \quad \text{and} \quad Y_{\bar{D}} = 1.5 + \langle\beta, X_{\bar{D}}\rangle + \varepsilon_{\bar{D}}$$

- *Scenario 2*

$$Y_D = 1 + 0.5\langle\beta, X^2_D\rangle + 2\varepsilon_D, \quad \text{and} \quad Y_{\bar{D}} = \langle\beta, X_{\bar{D}}\rangle + 1.5\varepsilon_{\bar{D}}$$

In both scenarios X_D and $X_{\bar{D}}$ were independently generated using (4.7) and X^2_D was obtained as the square of each element $X_D(t), 0 < t < 1$. Bearing in mind the distribution of the errors, ε_D and $\varepsilon_{\bar{D}}$, we have considered standard normal, Student-*t*, and skew normal distributions. In the functional setting, it is not straightforward identifying the covariate to induce the ROC curve. We have considered covariates of the form

$$X_z(t) = X_0(t) + zv_1(t),$$

where z lies between $(-2\sqrt{\lambda_1}, 2\sqrt{\lambda_1})$ and $(-1.5\sqrt{\lambda_1}, 1.5\sqrt{\lambda_1})$ in the simulation Scenarios 1 and 2, respectively. By varying z, we cover a wide range of possible covariates. Figure 2(b) in supplementary material available at *Biostatistics* online gives an idea of their shape.

To perform the computations, we determined the number of principal components retained by the PCV criterion in (2.4), and the bandwidth needed for the nonparametric model was chosen by the CV criterion in (2.6). We also need to specify the asymmetrical kernel as well as the most appropriate semimetric. The asymmetrical Gaussian kernel, $k(t) = \sqrt{2/\pi}\,\exp(-t^2/2)$ for $t \in (0, \infty)$, was used and the class of semimetrics $\{d_q^{\text{derivative}}\}_{q=0}^{2}$ was applied , where $d_q^{\text{derivative}}$denotes the L_2-norm of the qth derivative of the curve.

The discrepancy between the estimator and the true ROC curve is measured in terms of the empirical version of the global mean squared error (MSE)

$$\text{MSE} = \frac{1}{n_{X_z}} \sum_{l=1}^{n_{X_z}} \frac{1}{n_p} \sum_{r=1}^{n_p} (\widehat{\text{ROC}}_{X_{z_l}}(p_r) - \text{ROC}_{X_{z_l}}(p_r))^2.$$

The results of the simulation study pertaining to normally distributed errors are shown below; the results for the remaining distributions as well as further details and comparisons can be found in the supplementary material available at *Biostatistics* online.

The results are based on 1000 repetitions and, in all cases, the same sample size was considered, with $n_{\bar{D}} = n_D = 50, 100, 200$. Under Scenario 1, Figures 4 and 5, in supplementary material available at *Biostatistics* online, show the true ROC curve along with 2.5% and 97.5% simulation quantiles, for $z = -1.25, 0$, and 2, for the semiparametric and nonparametric estimators, respectively. The covariates corresponding to these z values are presented in Fig. 3 of the supplementary material available at *Biostatistics* online. As can be seen, although the semiparametric estimator displayed the lowest variance, both estimators recover the functional form of the true ROCs successfully. This can also be seen in Fig. 1. As expected, the variance of the estimates decreases as sample size increases. For Scenario 2, where the covariate effect was far from linear, the estimates obtained by the semiparametric model were clearly unsuitable, as can be checked both from the covariate-specific ROCs (Fig. 6 in supplementary material available at *Biostatistics* online) and from the AUC curve (top of Fig. 2). In turn, the good performance of the nonparametric estimator is evident, with it recovering the functional form of the true ROCs (Fig. 7 in supplementary material available at *Biostatistics* online) and the true AUC successfully (bottom of Fig. 2).

Table 1 in supplementary material available at *Biostatistics* online summarizes the mean-squared error for each approach and scenario. In Scenario 1, the errors produced by the nonparametric approach are larger than the errors from the semiparametric approach. On the other hand, in Scenario 2, the errors produced by the semiparametric approach are much larger than the ones produced by the nonparametric approach. The mean-squared error decreases as sample size gets larger. Figure 8 of supplementary material available at *Biostatistics* online presents boxplots of the mean-squared errors produced by the two approaches for the different sample sizes considered. Table 4, also in the supplementary material available at *Biostatistics* online, presents the MSEs for the aforementioned z values; the same conclusions apply.

Regarding the performance of the two estimators under Student-t and skew normal errors, we found out that methods are robust to departures from normality; the results with skew normal distributed errors are competitive with the results of normal errors. With Student-t distributed errors, the difference is a little bit more pronounced but still give good results. From our simulation study, we can conclude that

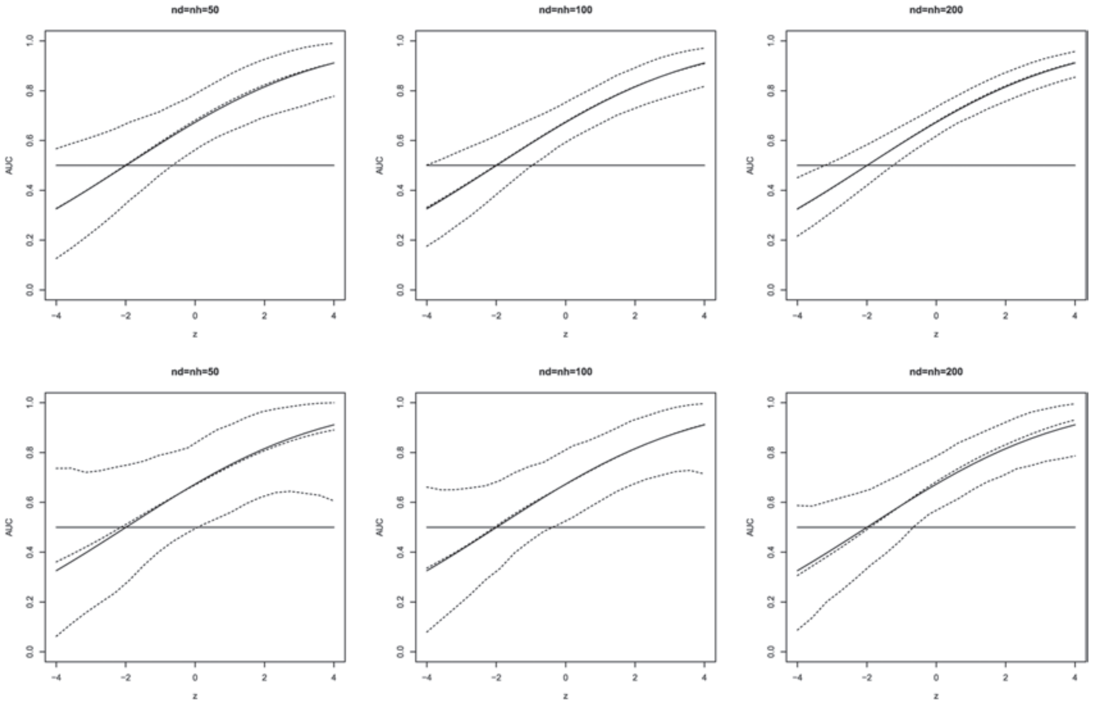

Fig. 1. True AUC (solid line) versus the average of simulated AUCs along with 2.5% and 97.5% simulation quantiles (dashed line) for Scenario 1. Top: semiparametric approach. Bottom: nonparametric approach.

the correct modeling of the covariate effect is far more important than the error distribution. A detailed summary of all these results can be found in the supplementary material available at *Biostatistics* online.

5. Application: metabolic syndrome study

Metabolic syndrome describes a cluster of abnormalities characterized by insulin resistance along with specific risk factors, including visceral adiposity, dyslipidaemia, and high blood pressure (Despres and Lemieux, 2006). Individuals with metabolic syndrome are at increased risk for cardiovascular disease (Lakka *and others*, 2002).

Serum GGT is a well-known marker of alcohol consumption and liver dysfunction. GGT is also associated with components of metabolic syndrome. Baseline serum GGT concentration appears to be an independent risk factor for the development of metabolic syndrome and the occurrence of cardiovascular disease and death (Lee *and others*, 2007). A hypothesis that appears to be consistent with these findings is that elevations of GGT are a marker of the presence of the metabolic syndrome. An important practical issue is whether, GGTs predictive properties can be used to identify people at high risk so that intervention to improve outcomes can be initiated. It is certainly easy and cheap to measure. In a recent study, Gude *and others* (2009) found, however, that serum concentrations of GGT are strongly associated with markers of nocturnal hypoxemia, particularly with arterial oxygen saturation levels during sleep. Since some conditions such as sleep disordering breathing or chronic obstructive pulmonary disease are very prevalent, it is important to study whether or not the performance of GGT at diagnosing metabolic syndrome may vary at different levels of arterial oxygen saturation.

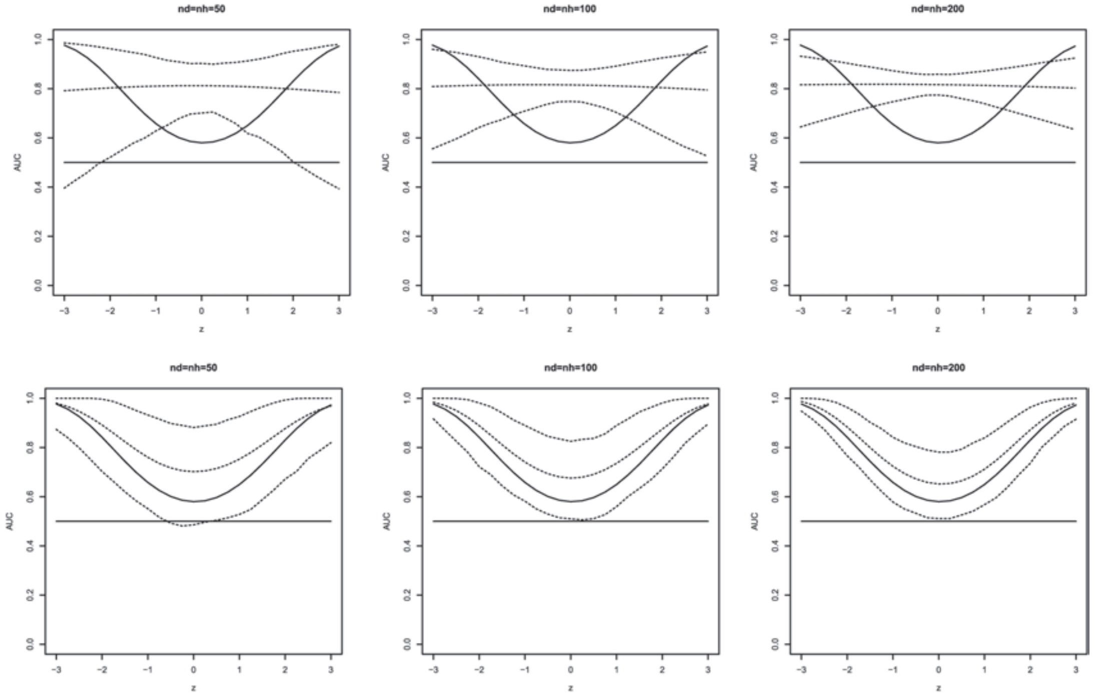

Fig. 2. True AUC (solid line) versus the average of simulated AUCs along with 2.5% and 97.5% simulation quantiles (dashed line) for Scenario 2. Top: semiparametric approach. Bottom: nonparametric approach.

With the aim of investigating this possible relationship, a study was conducted using a sample of 220 individuals. The present study took advantage of a survey of the general adult population from the municipality of A-Estrada, in northwestern Spain; detailed descriptions of study methodology and population sample characteristics have been reported elsewhere (González-Quintela *and others*, 2003; Gude *and others*, 2009).

The recording of arterial oxygen saturation was performed at the patient's home using a pulse oximeter with a finger probe. The arterial oxygen saturation was measured every 20 s thus leading to genuine functional data. As it is known that nocturnal oxygen arterial saturation has different patterns during the several sleep phases, for all individuals, we skipped the first 2 h of measurements and saved the next 3 h. Hence, at the final, we had a total of 540 measurements. Since GGT values are elevated among regular drinkers, we restricted the analysis to 115 women who reported no alcohol consumption. Thus, possible higher levels of GGT were not due to alcohol consumption and differences between genders. In short, the data analyzed here consist of 35 diseased and 80 healthy women. Our purpose is to investigate how the collected samples of arterial oxygen saturation affect the ability of GGT to accurately detect metabolic syndrome. The data analysis is divided into 2 parts. First, we examined the GGT performance as a marker to diagnose metabolic syndrome and then we conducted our functional ROC analysis that takes into account the effect of arterial oxygen saturation.

5.1 *ROC analysis of GGT discriminatory's capacity*

We carried out an initial analysis to evaluate the discriminatory capacity of the GGT in women, ignoring the arterial oxygen saturation effect. Here and in the subsequent analysis, we used the log-transformed

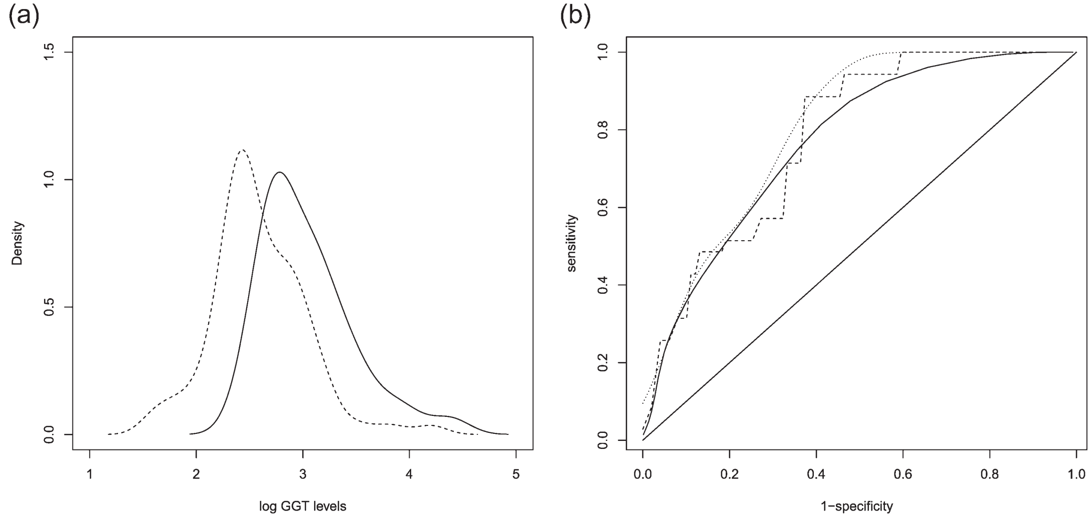

Fig. 3. (a) Densities of the log-transformed GGT levels in healthy (dashed) and diseased populations (solid); (b) ROC curve of log-transformed GGT measurements with no arterial oxygen saturation effect. The dashed curve is the empirical estimator, the solid line is the estimator proposed by Zou *and others* (1997) and the dotted curve is the estimator of Peng and Zhou (2004).

GGT levels. Figure 3(a) shows the densities of the log-transformed GGT levels in healthy and diseased populations, whereas Fig. 3(b) are the corresponding nonparametric estimators of the ROC curve. The curves lie well above the diagonal line, indicating a good discriminatory performance of GGT to distinguish between women with metabolic syndrome and those who are healthy. This can also be seen from the AUC, which is 0.773 (0.688, 0.857) for the empirical estimator and 0.769 (0.689, 0.846) and 0.801 (0.720, 0.869) for the Zou *and others* (1997) and Peng and Zhou (2004) estimators, respectively.

5.2 *Induced functional ROC regression analysis*

After analyzing the GGT's discriminatory capacity, we conducted our functional ROC analysis using the procedures described in the previous sections. It is known that the nocturnal arterial oxygen saturation shows different patterns between subjects with metabolic syndrome and healthy subjects (Gude *and others*, 2009). In Fig. 4 (top left and right), we can see clearly such difference. In the bottom panel of Fig. 4 is shown the smoothed mean and variance trajectories of the arterial oxygen saturation in each group (the smoothness of the curves was achieved with kernel smoothing). Once more, the different behavior of arterial oxygen saturation for healthy and diseased individuals is visible, with healthy subjects having higher arterial oxygen saturation levels and lowest variance.

We carried out a FPCs analysis for the predictor process by pooling together the 115 trajectories. The FPC scores displayed in Fig. 23(a) of the supplementary material available at *Biostatistics* online show a separation between diseased and healthy subjects. These first two components account for 81% of the variation in the data. Additionally, when we examined the plot of the log GGT levels against the estimated FPC scores in Fig. 23(b) of supplementary material available at *Biostatistics* online, the separation between the 2 groups still continue to be apparent.

To induce the covariate specific ROC curve, we have considered covariates of the form

$$X_z = \bar{X} + z\widehat{v}_1,$$

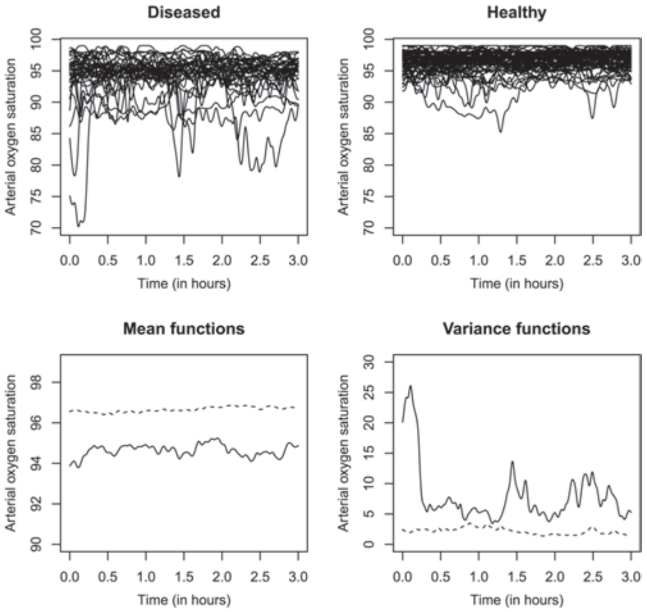

Fig. 4. Smoothed predictor trajectories in the diseased group (top left) and in the healthy group (top right). Mean (bottom left) and variance (bottom right) predictor trajectories. Solid and dashed lines correspond, respectively, to the diseased and healthy groups.

where $\bar{X}$ is the mean trajectory function of the pooled data, z is a weight parameter, and $\widehat{v}_1$ is the estimated eigenfunction associated with the first principal component. By inspection of Fig. 23(a) in supplementary material available at *Biostatistics* online, z was chosen to lie in the interval $(-50, 50)$. By varying z, we cover a wide range of covariate values and we therefore can investigate those covariates for which the marker is useful. Figure 24(a) of supplementary material available at *Biostatistics* online shows the 50 covariates, we have used to induce the ROC and Fig. 24(b), also in the supplementary material available at *Biostatistics* online, shows the covariates corresponding to z values of -40, 1, and 40. We remark that low values of z correspond to low values of arterial oxygen saturation, whereas higher values of z correspond to higher values of the oxygen saturation; as z increases, the arterial oxygen saturation curves also take higher values.

To determine whether the assumption of linearity is plausible for the data at hand, Chiou and Muller (2007) suggested to use the plots of the FPCs scores of the predictors curves against the response values. These authors have shown that if the model is correct, these plots may show linear relationships between the scores and the responses. Figure 25 in supplementary material available at *Biostatistics* online shows the plots for the first 5 principal components in the diseased group and the first 3 principal components in the healthy group, which appear to have a very slight relationship among some scores and the GGT values. To select the cutoff k_n, the PCV criterion in (2.4) was computed. Figure 5(a) shows the corresponding AUC curve. We have also included in the graph a confidence interval for the AUC, obtained by bootstrap. We use a bootstrap of residuals, which are real random variables, to resample the regression models and then the percentile method to obtain pointwise bootstrap confidence intervals for the AUC (500 bootstrap resamples). For further details, see González-Manteiga and Marínez-Calvo (2011) for the parametric case and Ferraty *and others* (2010) for the nonparametric one.

We then relaxed the linearity assumption and applied the nonparametric procedure. The smooth shape of the curves (top of Fig. 4) suggests to use the class of semimetrics $\{d_q^{\text{derivative}}\}_{q=0}^{2}$. We have also used

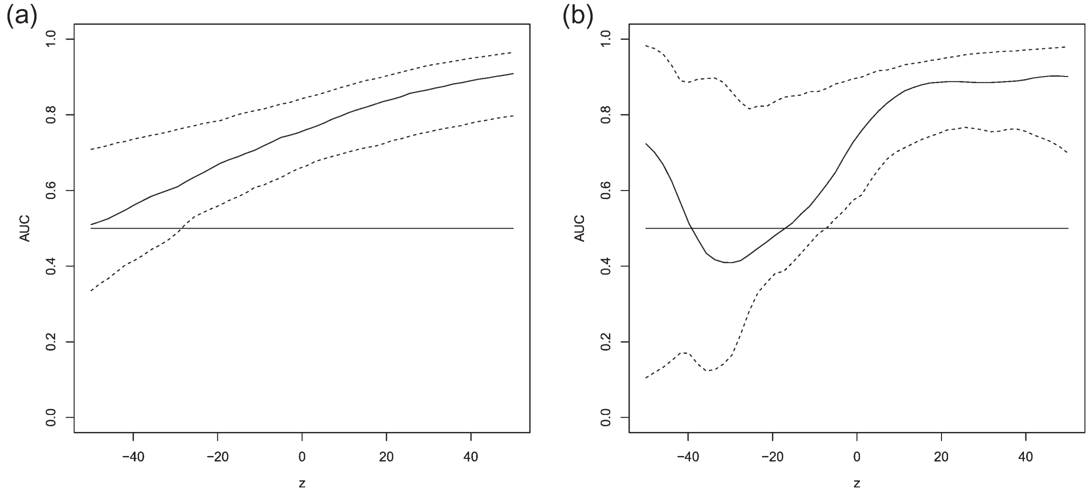

Fig. 5. AUC (solid line) together with the corresponding 95% bootstrap confidence bands (dashed lines). (a) Semiparametric functional model; (b) nonparametric functional model.

the asymmetrical Gaussian kernel and the bandwidth was determined in a data-driven way using (2.6). Figure 5(b) shows the corresponding estimates of AUC.

A large discrepancy between the 2 approaches is apparent. Based on the low values of the correlation between the FPC scores and the log GGT levels and based also on the results of our simulation study, this may indicate that the linearity assumption may not be valid for this data set. The nonparametric approach is more suitable due to its robustness, indicating that the arterial oxygen saturation affects the discriminatory capacity of the GGT as a marker to diagnose metabolic syndrome, with high values of oxygen saturation being associated with a better discrimination performance. In fact, Fig. 5(b) suggests that GGT has a better performance at the high values of oxygen saturation. These values are normal in healthy people (without chronic obstructive pulmonary disease or sleep apnea). At the low values, GGT has bad performance discriminating people with metabolic syndrome. At very low values, there are few cases, and the confidence intervals are wide. Thus, ignoring the oxygen saturation effect will result in an underestimated AUC for healthy subjects (those with high values of arterial oxygen saturation) and an overestimate for those suffering from chronic obstructive pulmonary disease and apnea (and hence with low values of oxygen saturation during night). These findings indicate that nocturnal hypoxemia should be taken into account when interpreting serum levels of GGT in clinical practice. Specifically, performance of GGT is good in "healthy" subjects but not in individuals with chronic obstructive pulmonary disease or sleep apnea.

6. Discussion

In this paper, we discuss the extension to the functional context of induced ROC methodology. The need for this modeling approach was motivated by a real-data example, where we evaluated the effect of arterial oxygen saturation (measured densely over night) on GGT's discriminatory capacity to detect metabolic syndrome in women. Semiparametric and nonparametric approaches were considered for estimating the induced functional ROC curve. Simulation results indicate a better performance of the semiparametric approach when the linearity assumption holds. On the other hand, the nonparametric approach—even with a loss of accuracy in the parametric case—overcomes the linearity issue, being thus more robust and flexible enough to model many practical situations. We thus provided a versatile class of models to estimate the induced ROC from which the data analyst can choose the most appropriate one for the data at hand. We point out that the methods are not computationally time consuming. `R` code is available from the first author on request. Extensions to the functional context of direct ROC methodology warrant future research.

Supplementary material

Supplementary material is available at http://biostatistics.oxfordjournals.org.

Acknowledgments

We thank the editor, the associate editor, and the 2 reviewers for their valuable comments and suggestions. Vanda Inácio also thanks Antónia Amaral Turkman and Miguel de Carvalho for helpful comments and support.

Funding

This research is funded by the Portuguese Foundation for Science and Technology through the PhD (SFRH/BD/47742/2008) and through the projects PEst-OE/MAT/UI0006/2011 and PTDC/MAT/118335/2010, and by the Spanish Ministry of Science and Innovation through the projects MTM2008-03010, MTM2008-01603, and MTM2011-28285-C02-01.

References

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free ROC analysis using binary regression techniques. *Biostatistics* **3**, 421–432.

Cai, T. (2004). Semiparametric ROC regression analysis with placement values. *Biostatistics* **5**, 45–60.

Cai, T. T. and Hall, P. (2006). Prediction in functional linear regression. *Annals of Statistics* **34**, 2159–2179.

Cardot, H., Ferraty, F. and Sarda, P. (1999). Functional linear model. *Statistics and Probability Letters* **45**, 11–22.

Chiou, J. M. and Muller, H. G. (2007). Diagnostics for functional regression via residual processes. *Computational Statistics and Data Analysis* **51**, 4849–4863.

Despres, J. P. and Lemieux, I. (2006). Abdominal obesity and metabolic syndrome. *Nature* **444**, 881–887.

Etzioni, R., Pepe, M. S., Longton, G., Hu, C. and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: a case study of prostate cancer. *Medical Decision Making* **19**, 242–251.

Farraggi, D. (2003). Adjusting receiver operating characteristic curves and related indices for covariates. *The Statistician* **52**, 179–192.

Febrero-Bande, M., Galeano, P. and González-Manteiga, W. (2010). Measures of influence for the functional linear model with scalar response. *Journal of Multivariate Analysis* **101**, 327–339.

Ferraty, F. and Vieu, P. (2002). The functional nonparametric model and application to spectrometric data. *Computational Statistics* **17**, 545–564.

Ferraty, F. and Vieu, P. (2006). *Nonparametric Functional Data Analysis.* New York: Springer.

Ferraty, F., Van Keilegom, I. and Vieu, P. (2010). On the validity of the bootstrap in non-parametric functional regression. *Scandinavian Journal of Statistics* **37**, 286–306.

González-Manteiga, W. and Martínez-Calvo, A. (2011). Bootstrap in functional linear regression. *Journal of Statistical Planning and Inference* **141**, 453–461.

González-Manteiga, W., Pardo-Fernandéz, J. C. and Van Keilegom, I. (2011). ROC curves in non-parametric location-scale regression models. *Scandinavian Journal of Statistics* **38**, 169–184.

González-Quintela, A., Gude, F., Boquete, O., Rey, J., Meijide, L. M., Suarez, F., Fernández-Merino, M. C., Pérez, L. F. and Vidal, C. (2003). Association of alcohol consumption with total serum immunoglobulin E levels and allergic sensitization in an adult population-based survey. *Clinical and Experimental Allergy* **33**, 199–205.

Gude, F., Rey-Garcia, J., Fernandez-Merino, C., Meijide, L., García-Ortiz, L., Zamarron, C. and Gonzalez-Quintela, A. (2009). Serum levels of gamma-glutamyl transferase are associated with markers of nocturnal hypoxemia in a general adult population. *Clinica Chimica Acta* **407**, 67–71.

Lakka, H. M., Laaksonen, D. E., Lakka, T. A., Niskonen, L. K., Kumpuselo, E., Tuomilehto, J. and Salonen, J. T. (2002). The metabolic syndrome and total and cardiovascular disease mortality in middle-aged men. *Journal of the American Medical Association* **288**, 2709–2716.

Lee, D. S., Evans, J. C., Robins, S. J., Wilson, P. W., Albano, I., Fox, C. S., Wang, T. J., Benjamin, E. J., Vasan, R. S. (2007). Gamma glutamyl transferase and metabolic syndrome, cardiovascular disease, and mortality risk: the Framingham Heart Study. *Arteriosclerosis, Trombosis, and Vascular Biology* **27**, 127–133.

Nadaraya, E. A. (1964). On estimating regression. *Theory of Probability and its Applications* **9**, 141–142.

Peng, L. and Zhou, X. H. (2004). Local linear smoothing of receiver operating characteristic (ROC) curves. *Journal of Statistical Planning and Inference* **118**, 129–143.

Pepe, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 124–135.

Pepe, M. S. (2003). *The Statistical Evaluation of Medical Tests for Classification and Prediction.* New York: Oxford University Press.

Ramsay, J. O. and Delzel, C. J. (1991). Some tools for functional data analysis. *Journal of the Royal Statistical Society, Series B* **53**, 539–572.

Ramsay, J. O. and Silverman, B. (2006). *Functional Data Analysis.* New York: Springer.

Rodríguez-Álvarez, M. X., Tahoces, P. C., Cadarso-Suárez, C. and Lado, M. J. (2011a). Comparative study of ROC regression techniques—applications for the computer-aided diagnostic system in breast cancer detection. *Computational Statistics and Data Analysis* **55**, 888–902.

Rodríguez-Álvarez, M. X., Roca-Pardiñas, J. and Cadarso-suárez, C. (2011b). ROC curve and covariates: extending the induced methodology to the non-parametric framework. *Statistics and Computing* **21**, 483–485.

Watson, G. S. (1964). Smooth regression analysis. *Sankhyã* **26**, 359–372.

Yao, F., and Müller, H. G. (2010). Functional quadratic regression. *Biometrika* **97**, 49–64.

Zhao, X., Marron, J. S. and Wells, M. T. (2004). The functional data analysis view of longitudinal data. *Statistica Sinica* **14**, 789–808.

Zheng, Y. and Heagerty, P. J. (2004). Semiparametric estimation of time dependent ROC curves for longitudinal marker data. *Biostatistics* **4**, 615–632.

Zou, K. H., Hall, W. J. and Shapiro, D.E. (1997). Smooth nonparametric receiver operating characteristic (ROC) curves for continuous diagnostic tests. *Statistics in Medicine* **16**, 2143–2156.

[*Received August 30, 2011; revised February 17, 2012; accepted for publication February 28, 2012*]