Computational Statistics and Data Analysis 55 (2011) 888–902

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Comparative study of ROC regression techniques—Applications for the computer-aided diagnostic system in breast cancer detection

María Xosé Rodríguez-Álvarez [a,b,c,*], Pablo G. Tahoces [d], Carmen Cadarso-Suárez [a,c], María José Lado [e]

[a] Unit of Biostatistics, Department of Statistics and Operations Research, University of Santiago de Compostela, Spain
[b] Complexo Hospitalario Universitario de Santiago de Compostela (CHUS), Santiago de Compostela, Spain
[c] Instituto de Investigación Sanitaria de Santiago de Compostela (IDIS), Santiago de Compostela, Spain
[d] Department of Electronics and Computer Science, University of Santiago de Compostela, Spain
[e] Department of Computer Science, University of Vigo, Spain

ARTICLE INFO

Article history:
Received 1 September 2009
Received in revised form 19 July 2010
Accepted 20 July 2010
Available online 29 July 2010

Keywords:
ROC curve
Regression techniques
B-splines
Computer-aided diagnosis

ABSTRACT

The receiver operating characteristic (ROC) curve is the most widely used measure for statistically evaluating the discriminatory capacity of continuous biomarkers. It is well known that, in certain circumstances, the markers' discriminatory capacity can be affected by factors, and several ROC regression methodologies have been proposed to incorporate covariates in the ROC framework. An in-depth simulation study of different ROC regression models and their application in the emerging field of automatic detection of tumour masses is presented. In the simulation study different scenarios were considered and the models were compared to each other on the basis of the mean squared error criterion. The application of the reviewed ROC regression techniques in evaluating computer-aided diagnostic (CAD) schemes can become a major factor in the development of such systems in Radiology.

1. Introduction

The statistical evaluation of the performance of a continuous marker, Y, in distinguishing healthy individuals from diseased individuals is usually based on the receiver operating characteristic (ROC) curve (Metz, 1978). Under the conventional assumption that high marker values are indicative of disease, classification of an individual as healthy ($\bar{D}$) or diseased (D) on the basis of Y can be made by the choice of a cut-off value c, such that if $Y \geq c$, the individual is classified as diseased and if $Y < c$, the individual is classified as healthy. Hence, each chosen cut-off value c will give rise to a true positive fraction, $\mathrm{TPF}(c) = P[Y \geq c|D]$, and a false positive fraction, $\mathrm{FPF}(c) = P[Y \geq c|\bar{D}]$. In such a situation, the ROC curve is defined as the set of all TPF–FPF pairs that can be obtained with varying cut-off values for c, $\{(\mathrm{FPF}(c), \mathrm{TPF}(c)), c \in (-\infty, \infty)\}$. In practice, the ROC curve is usually represented as a function of the form

$$ROC(t) = S_D\left(S_{\bar{D}}^{-1}(t)\right), \quad t \in (0, 1),$$

where $S_D(y) = P[Y \geq c|D]$ and $S_{\bar{D}}(y) = P[Y \geq c|\bar{D}]$ denote the survival functions of Y in diseased and healthy subjects, respectively.

* Corresponding address: Unit of Biostatistics, Faculty of Medicine, San Francisco s/n., 15782, Santiago de Compostela, Spain. Tel.: +34 981 563100x12281; fax: +34 981 547172.
E-mail address: mariajose.rodriguez.alvarez@usc.es (M.X. Rodríguez-Álvarez).

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.07.018

In many practical situations, however, a marker's discriminatory capacity may be affected by covariates (such as, e.g., the age or gender of the individual). In such situations, the ROC curve (or other summary indices, such as the area under the ROC curve, AUC) may be of little value if important covariates are ignored, and the interest must be focused on assessing marker Y's discriminatory capacity by reference to the values assumed by the vector of covariates $\boldsymbol{X}$. If the conditional survival functions of Y_D and $Y_{\bar{D}}$ given $\boldsymbol{X}$ are denoted by $S_{D\boldsymbol{X}}$ and $S_{\bar{D}\boldsymbol{X}}$ respectively, the conditional or covariate-specific ROC curve is defined as

$$ROC_{\boldsymbol{X}}(t) = S_{D\boldsymbol{X}}\left(S^{-1}_{\bar{D}\boldsymbol{X}}(t)\right), \quad t \in (0, 1). \tag{1}$$

Various statistical methods have been proposed in the literature to incorporate covariate information in the ROC-based analysis. In this paper, the focus is on two methodologies that can be viewed as coming within the general framework of regression, namely: (1) 'induced' methodology (Faraggi, 2003; Pepe, 1998; Tosteson and Begg, 1988; Zheng and Heagerty, 2004) and (2) 'direct' methodology (Alonzo and Pepe, 2002; Cai, 2004; Cai and Pepe, 2002; Pepe, 2000). 'Induced' methodology is based on using separate regression models for the marker in healthy, $Y_{\bar{D}}$, and diseased, Y_D, populations respectively. Covariate effects on the associated ROC curve can then be computed by deriving the induced form of the ROC curve. Instead of targeting the marker, 'direct' methodology assumes a regression model for the conditional ROC curve, $ROC_{\boldsymbol{X}}(t)$, with the effect of the covariates thus being directly evaluated on the ROC curve.

To date, only partial comparative simulation studies of direct and induced methodologies have been published (Cai, 2004; Cai and Pepe, 2002; Pepe, 1998). Accordingly, one of the main goals of this paper is to perform an in-depth simulation study to (a) compare statistically different ROC regression approaches and (b) evaluate their robustness in the face of deviations from the 'theoretical' model.

With respect to induced methodology, in this study the theoretical distribution of the ROC regression parameters in the case of Gaussian errors is presented. Moreover, where no assumption is made about the distributions of the errors, Pepe (1998)'s study is extended to allow for different distributions in diseased and healthy subjects.

This paper also studies the benefits of applying these ROC regression methodologies in the emerging field of automatic detection of tumour masses in breast cancer. The inclusion of covariates in ROC analysis within the framework of computer-aided diagnosis (CAD) systems has recently been addressed by López de Ullibarri et al. (2008a). That paper proposes an estimator of the covariate-specific ROC curve based on local linear estimation of the conditional survival functions in healthy and diseased subjects. However, the authors do not propose inference procedures to examine the effects of covariates on the ROC curve. This study applies to a CAD system the reviewed ROC regression approaches that might represent a useful alternatives to including covariates in the ROC curve, and that would enable inferences to be made about the effects of such covariates.

So far, the scarcity of implemented ROC regression software is probably responsible for these models' lack of popularity in the medical community. Furthermore, only the Alonzo and Pepe (2002)'s approach is implemented in the commercial statistical software Stata (Janes and Pepe, 2009). Therefore, a user-friendly R package (R Development Core Team, 2010), known as ROC Regression, was also developed (López de Ullibarri et al., 2008b). This software provides numerical and graphical outputs of the different ROC regression methods reviewed in this study.

The layout of this paper is as follows. Section 2 presents induced and direct ROC regression methodologies. The results of the simulation study are reported in Section 3. In Section 4, ROC regression techniques are applied to a CAD system dedicated to the detection of malignant masses in digital mammograms. Lastly, the paper concludes with a discussion in Section 5.

2. ROC regression methods

This section describes direct and induced ROC regression methodologies, together with the estimation procedures in each case. It should be pointed out that, though the models are going to be presented, for the sake of simplicity, under the assumption of linearity in the effects of the continuous variables, both methodologies allow for incorporation of flexibility in the modelling of these effects, by the use of, say, regression splines (de Boor, 2001). Indeed, this situation was also considered both in the simulation study and in application to CAD system data.

Lastly, for study purposes deem $\boldsymbol{X}$ to be a p-dimensional covariate vector and let $\{(y_{\bar{D}i}, \boldsymbol{x}_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_{Dj}, \boldsymbol{x}_{Dj})\}_{j=1}^{n_D}$ be two independent random samples drawn from the healthy and diseased populations respectively.

2.1. Induced ROC regression methodology

Induced ROC methodology is based on modelling the result of the marker for diseased and healthy subjects respectively, according to the covariate vector $\boldsymbol{X}$, so that, on the basis of the two models, the expression of the covariate-specific ROC curve can be induced. Based on our application (see Section 4 for details), we focus our attention on homoscedastic regression models. More specifically, let

$$Y_D = \widetilde{\boldsymbol{X}}'\boldsymbol{\beta}_D + \sigma_D\varepsilon_D \tag{2a}$$

and

$$Y_{\bar{D}} = \widetilde{\boldsymbol{X}}'\boldsymbol{\beta}_{\bar{D}} + \sigma_{\bar{D}}\varepsilon_{\bar{D}}, \tag{2b}$$

where $\widetilde{\boldsymbol{X}} = (1, \boldsymbol{X}')'$, $\boldsymbol{\beta}_D = (\beta_{D0}, \beta_{D1}, \ldots, \beta_{Dp})'$ and $\boldsymbol{\beta}_{\bar{D}} = (\beta_{\bar{D}0}, \beta_{\bar{D}1}, \ldots, \beta_{\bar{D}p})'$ are $(p+1)$-dimensional vectors of unknown parameters, and ε_D and $\varepsilon_{\bar{D}}$ are random independent variables having a mean of zero, a variance of one, and survival functions S_D and $S_{\bar{D}}$ respectively. Based on (1) and (2), the expression of the covariate-specific ROC curve is easily obtained:

$$ROC_{\boldsymbol{X}}(t) = S_D\left(\widetilde{\boldsymbol{X}}'\boldsymbol{\beta} + \alpha S_{\bar{D}}^{-1}(t)\right), \quad t \in (0, 1), \tag{3}$$

where $\boldsymbol{\beta} = \frac{\boldsymbol{\beta}_{\bar{D}} - \boldsymbol{\beta}_D}{\sigma_D}$ is a $(p+1)$-dimensional vector, $\alpha = \frac{\sigma_{\bar{D}}}{\sigma_D}$ and $S_{\bar{D}}^{-1}(t) = \inf\{y | S_{\bar{D}}(y) \leq t\}$.

As can be seen in expression (3), in induced methodology the effect of a covariate on the ROC curve can be quantified on the basis of the difference of the effects (in quantitative terms) of the covariate on the healthy and diseased populations. Accordingly, where its effect on healthy and diseased subjects coincides, a covariate will have no effect on a marker's discriminatory capacity.

This study reviews two models that differ in their assumptions about the distribution of errors ε_D and $\varepsilon_{\bar{D}}$ in (2), namely: (a) the normal model (Faraggi, 2003), which assumes Gaussian errors; and (b) the semiparametric model (Pepe, 1998), where the distributions of the errors are not specified.

2.1.1. Induced normal model (I1)

Under the assumption of normality of errors ε_D and $\varepsilon_{\bar{D}}$ in (2), the covariate-specific ROC curve (3) follows the so-called binormal model. In such a case, the estimation procedure consists of the following steps:

1. estimate $\boldsymbol{\beta}_{\bar{D}}$ and $\boldsymbol{\beta}_D$ by ordinary least squares, on the basis of samples $\{(y_{\bar{D}i}, \boldsymbol{x}_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$ and $\{(y_{Dj}, \boldsymbol{x}_{Dj})\}_{j=1}^{n_D}$ respectively;
2. estimate σ_D^2 and $\sigma_{\bar{D}}^2$ as

$$\hat{\sigma}_D^2 = \frac{\sum_{j=1}^{n_D}\left(y_{Dj} - \widetilde{\boldsymbol{x}}_{Dj}'\hat{\boldsymbol{\beta}}_D\right)^2}{n_D - p - 1} \quad \text{and} \quad \hat{\sigma}_{\bar{D}}^2 = \frac{\sum_{i=1}^{n_{\bar{D}}}\left(y_{\bar{D}i} - \widetilde{\boldsymbol{x}}_{\bar{D}i}'\hat{\boldsymbol{\beta}}_{\bar{D}}\right)^2}{n_{\bar{D}} - p - 1};$$

and,

3. finally, calculate the covariate-specific ROC curve as follows

$$\widehat{\mathrm{ROC}}_X(t) = \Phi\left(-\widetilde{\boldsymbol{X}}'\hat{\boldsymbol{\beta}} + \hat{\alpha}\Phi^{-1}(t)\right), \quad t \in (0, 1),$$

where Φ is cumulative distribution function of a standard normal random variable.

Given that the covariate-specific ROC curve is a binormal ROC curve, for this model the conditional AUC, $AUC_{\boldsymbol{X}} = \int_0^1 ROC_{\boldsymbol{X}}(t)\mathrm{d}t$, has a closed form. Its expression and the estimator, along with the procedure for constructing approximate $100 \times (1-\alpha)$ per cent confidence intervals, can be found in Faraggi (2003).

In the present study, we derive the distribution of the induced ROC parameter estimates $\hat{\boldsymbol{\beta}} = (\hat{\boldsymbol{\beta}}_{\bar{D}} - \hat{\boldsymbol{\beta}}_D)/\hat{\sigma}_D$. From the standard linear model theory, one has

$$\hat{\boldsymbol{\beta}}_D \sim N(\boldsymbol{\beta}_D, \Sigma_D^2), \qquad \hat{\boldsymbol{\beta}}_{\bar{D}} \sim N(\boldsymbol{\beta}_{\bar{D}}, \Sigma_{\bar{D}}^2) \quad \text{and} \quad (n_D - p - 1)\,\frac{\hat{\sigma}_D^2}{\sigma_D^2} \sim \chi^2_{n_D - p - 1},$$

where $\Sigma_D^2 = \sigma_D^2(\widetilde{\boldsymbol{x}}_D\widetilde{\boldsymbol{x}}_D')^{-1}$ and $\Sigma_{\bar{D}}^2 = \sigma_{\bar{D}}^2(\widetilde{\boldsymbol{x}}_{\bar{D}}\widetilde{\boldsymbol{x}}_{\bar{D}}')^{-1}$, with $\hat{\boldsymbol{\beta}}_D$, $\hat{\boldsymbol{\beta}}_{\bar{D}}$ and $\hat{\sigma}_D^2$ being independently distributed.

Performing some simple algebraic operations, it can thus be shown that

$$M * \hat{\boldsymbol{\beta}} \sim t\left(M * \boldsymbol{\beta}, \frac{1}{\sigma_D^2}MM' * (\Sigma_D^2 + \Sigma_{\bar{D}}^2), n_D - p - 1\right),$$

where the operator $*$ represents the element by element arithmetic multiplication, t denotes de multivariate non-central t-distribution (Kshirsagar, 1961), $M = \sigma_D((\sigma^2_{\hat{\beta}_{D0}} + \sigma^2_{\hat{\beta}_{\bar{D}0}})^{-1/2}, \ldots, (\sigma^2_{\hat{\beta}_{Dp}} + \sigma^2_{\hat{\beta}_{\bar{D}p}})^{-1/2})'$, and $\sigma_{\hat{\beta}_{Di}}$ and $\sigma_{\hat{\beta}_{\bar{D}i}}$ are the variances of $\hat{\beta}_{Di}$ and $\hat{\beta}_{\bar{D}i}$ respectively. Accordingly, the marginal distribution of any of the components of $M * \hat{\boldsymbol{\beta}}$ will be the non-central t-distribution, namely

$$M_i\hat{\beta}_i \sim t\,(M_i\beta_i, n_D - p - 1), \quad i = 0, \ldots, p,$$

where t denotes de univariate non-central t-distribution, and M_i is the i-element of vector M. Hence, $100 \times (1-\alpha)$ per cent approximate confidence limits for $\hat{\beta}_i$ can be obtained through limits for the non-centrality parameter (Venables, 1975), by replacing M_i by its empirical counterpart, $\hat{M}_i = \frac{\hat{\sigma}_D}{\sqrt{\hat{\sigma}^2_{\hat{\beta}_{Di}} + \hat{\sigma}^2_{\hat{\beta}_{\bar{D}i}}}}$, where $\hat{\sigma}^2_{\hat{\beta}_{Di}} = \hat{\sigma}_D^2(\widetilde{\boldsymbol{x}}_D\widetilde{\boldsymbol{x}}_D')^{-1}_{ii}$ and $\hat{\sigma}^2_{\hat{\beta}_{\bar{D}i}} = \hat{\sigma}_{\bar{D}}^2(\widetilde{\boldsymbol{x}}_{\bar{D}}\widetilde{\boldsymbol{x}}_{\bar{D}}')^{-1}_{ii}$.

2.1.2. Induced semiparametric model (I2)

In contrast to Pepe (1998), where the same distribution was assumed for errors ε_D and $\varepsilon_{\bar{D}}$ in (2), different distributions will be considered here.

As in the previous case, in this model estimation of the covariate-specific ROC curve (3) firstly entails estimating the vectors of unknown parameters $\boldsymbol{\beta}_D$ and $\boldsymbol{\beta}_{\bar{D}}$, as well as the variances σ_D^2 and $\sigma_{\bar{D}}^2$ in (2). For the case in point, this estimation may be made in the same way as that shown for the model I1 (steps 1 and 2 of the previous algorithm), since the estimation of parameters $\boldsymbol{\beta}_D$ and $\boldsymbol{\beta}_{\bar{D}}$ using ordinary least squares coincides with the estimation using quasi-likelihood-based methods (McCullagh and Nelder, 1989). Once these values have been estimated, the remaining steps for estimating the conditional ROC curve would be as follows:

3. estimate survival functions S_D and $S_{\bar{D}}$ on the basis of the empirical distribution of the standardised residuals

$$\hat{S}_D(y) = \frac{1}{n_D}\sum_{j=1}^{n_D} I\left[\frac{y_{Dj} - \widetilde{\boldsymbol{x}}'_{Dj}\hat{\boldsymbol{\beta}}_D}{\hat{\sigma}_D} \geq y\right]$$

and

$$\hat{S}_{\bar{D}}(y) = \frac{1}{n_{\bar{D}}}\sum_{i=1}^{n_{\bar{D}}} I\left[\frac{y_{\bar{D}i} - \widetilde{\boldsymbol{x}}'_{\bar{D}i}\hat{\boldsymbol{\beta}}_{\bar{D}}}{\hat{\sigma}_{\bar{D}}} \geq y\right].$$

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

$$\widehat{\mathrm{ROC}}_X(t) = \hat{S}_D\left(\widetilde{\boldsymbol{X}}'\hat{\boldsymbol{\beta}} + \hat{\alpha}\hat{S}_{\bar{D}}^{-1}(t)\right), \quad t \in (0, 1),$$

where $\hat{S}_{\bar{D}}^{-1}(t) = \inf\{y|\hat{S}_{\bar{D}}(y) \leq t\}$.

Insofar as the inference is concerned, bootstrapping techniques (Efron and Tibshirani, 1993) can be used to compute the standard errors of the induced ROC parameter estimates $\hat{\boldsymbol{\beta}}$.

2.2. Direct ROC regression methodology

In contrast to induced methodology, in direct methodology the effect of the covariates is directly evaluated on the ROC curve, within the framework of the generalised linear model (GLM) (McCullagh and Nelder, 1989). In this methodology, the general form of the ROC curve is given by the following regression model

$$ROC_{\boldsymbol{X}}(t) = g\left(h_0(t) + \boldsymbol{X}'\boldsymbol{\beta}\right), \quad t \in (0, 1), \tag{4}$$

where $g(\cdot)$ is a known function (the inverse of the link function), $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)'$ is a p-dimensional vector of unknown parameters, and $h_0(\cdot)$ is an unknown monotonic increasing function. Models like (4) define the so-called class of ROC–GLMs (Pepe, 2003).

Unlike standard regression analysis, in ROC–GLM regression models (4) the dependent variable is not directly observable. This makes it necessary for another interpretation to be given to the ROC curve (Pepe, 2000; Pepe and Cai, 2004). The key idea for fitting the model (4) is based on the placement values (Hanley and Hajian-Tilaki, 1997) of Y_D, defined as $PV_D \equiv S_{\bar{D}\boldsymbol{X}}(Y_D)$. Given that,

$$\begin{aligned} P[PV_D \leq t|\boldsymbol{X}] &= P[S_{\bar{D}\boldsymbol{X}}(Y_D) \leq t|\boldsymbol{X}] = P[Y_D \geq S_{\bar{D}\boldsymbol{X}}^{-1}(t)|\boldsymbol{X}] \\ &= S_{D\boldsymbol{X}}(S_{\bar{D}\boldsymbol{X}}^{-1}(t)) \\ &= ROC_{\boldsymbol{X}}(t), \end{aligned} \tag{5}$$

the covariate-specific ROC curve can be viewed in two different ways: (a) as the conditional expectation of binary variable $B_{Dt} = I[PV_D \leq t]$; or, (b) as the cumulative distribution function of placement values, PV_D.

Shown below are two approaches to direct ROC regression methodology, which differ in the assumptions made about function h_0 in (4).

2.2.1. Direct parametric ROC–GLM (D1)

In parametric ROC–GLM methodology (Alonzo and Pepe, 2002; Pepe, 2000), a parametric form for the baseline function $h_0(\cdot)$ of the model (4) is specified, $h_0(t) = \sum_{k=1}^{K}\alpha_k h_k(t)$ where $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$ is a vector of unknown parameters and $h(\cdot) = (h_1(\cdot), \ldots, h_K(\cdot))$ are known functions. This configuration therefore gives rise to a completely parametric ROC–GLM model (4):

$$ROC_{\boldsymbol{X}}(t) = g\left(\sum_{k=1}^{K}\alpha_k h_k(t) + \boldsymbol{X}'\boldsymbol{\beta}\right), \quad t \in (0, 1). \tag{6}$$

Based on the interpretation of the covariate-specific ROC curve as the conditional expectation of binary variable $B_{Dt} = I[PV_D \leq t]$ (see (5)), the ROC–GLM regression model (6) can be viewed as a regression model for B_{Dt}. This suggests that estimation of the parametric ROC–GLM regression model (6) can be based on the following algorithm:

1. choose a set $T = \{t_l : l = 1, \ldots, n_T\} \subset (0, 1)$ of FPF's;
2. estimate $S_{\bar{D}\boldsymbol{X}}$ on the basis of $\{(y_{\bar{D}i}, \boldsymbol{x}_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$;
3. for each disease observation, calculate the estimated placement values $PV_j = \hat{S}_{\bar{D}\boldsymbol{x}_{Dj}}(y_{Dj}), j = 1, \ldots, n_D$;
4. for all $t \in T$ and each disease observation, calculate the binary placement value indicator $\hat{B}_{jt} = I[PV_j \leq t], t \in T$, $j = 1, \ldots, n_D$; and,
5. fit the marginal binary regression ROC–GLM model (6) to the data $\{(\hat{B}_{jt}, \{\boldsymbol{x}_{Dj}, h_1(t), \ldots, h_k(t)\}), t \in T, j = 1, \ldots, n_D\}$.

It should be pointed out that the above estimation procedure makes it possible for observations to be clustered, e.g., when individuals are tested several times. With respect to the inference, in the case where observations are independent, the asymptotic distribution of the parameter estimates has been stated (Pepe, 2000). In practice, however, Alonzo and Pepe (2002) suggest the use of bootstrap methods. In such a case, the data should be resampled in accordance with the study design and presence of cluster data.

2.2.2. Direct semiparametric ROC–GLM (D2)

In contrast to model D1 described above, in the semiparametric ROC–GLM (Cai, 2004; Cai and Pepe, 2002) no assumptions are made about function $h_0(\cdot)$ in (4), which is left unspecified. In this approach, the key idea is that the ROC–GLM regression model (4) can be viewed as a regression model for placement values PV_D. It is easy to see that regression model (4) is equivalent to

$$h_0(PV_D) = -\boldsymbol{X}'\boldsymbol{\beta} + \varepsilon, \tag{7}$$

where ε is a random variable with known distribution g, and $h_0(\cdot)$ is a completely unspecified increasing function. This equivalence leads to the idea of using pairwise comparison of placement values for estimating the regression parameters $\boldsymbol{\beta}$ of (7) (Cai, 2004; Cheng et al., 1995). In the estimation procedure of the semiparametric ROC–GLM model, the effect of the covariates on the ROC curve are firstly estimated, by estimating the vector of unknown parameters $\boldsymbol{\beta}$. Finally, the previous estimations are used to estimate the baseline function $h_0(\cdot)$.

For the sake of simplicity, observations among the healthy and diseased populations will be assumed to be independent. Nevertheless, with certain amendments, the following estimation procedure can incorporate cluster data.

Estimation of the vector of unknown parameters $\boldsymbol{\beta}$

In this case, the steps of the algorithm are as follows:

1. estimate $S_{\bar{D}\boldsymbol{X}}$ on the basis of $\{(y_{\bar{D}i}, \boldsymbol{x}_{\bar{D}i})\}_{i=1}^{n_{\bar{D}}}$;
2. for each disease observation, calculate the estimated placement values $PV_j = \hat{S}_{\bar{D}\boldsymbol{x}_{Dj}}(y_{Dj}), j = 1, \ldots, n_D$;
3. for each pair of observations in the diseased population, calculate $\widehat{PV}_{jl}^{W} = I[PV_j \leq PV_l]$ and $\boldsymbol{x}_{jl} = \boldsymbol{x}_{Dj} - \boldsymbol{x}_{Dl}$ with $j, l = 1, \ldots, n_D, j \neq l$; and,
4. based on data $\{(\widehat{PV}_{jl}^{W}, \boldsymbol{x}_{jl}), j, l = 1, \ldots, n_D, j \neq l\}$ estimate the ROC regression parameters $\boldsymbol{\beta}$ by solving

$$\sum_{j=1}^{n_D} \sum_{l<j} \omega_\xi(\boldsymbol{x}_{jl}'\boldsymbol{\beta})\boldsymbol{x}_{jl} \left\{ \widehat{PV}_{jl}^{W} - \xi(\boldsymbol{x}_{jl}'\boldsymbol{\beta}) \right\} = 0,$$

where $\omega_\xi(x) = \frac{\mathrm{d}\xi(x)/\mathrm{d}x}{\xi(x)(1-\xi(x))}$ and $\xi(x) = \int_{-\infty}^{-\infty} g(x+y)\mathrm{d}y$. It should be noted that the above estimation function assumes that the binary variables $\{\widehat{PV}_{jl}^{W}, j, l = 1, \ldots, n_D, j \neq l\}$ are independent.

Like model D1, in order to put the above algorithm into practice an estimator for the conditional survival function must be chosen in step 1. Insofar as step 4 is concerned, the solution to the estimation equation entails choosing the link function g^{-1} (some popular choices are, e.g., probit or logit).

As regards the inference, Cai (2004) shows the consistency and asymptotic distribution of the parameter estimator $\hat{\boldsymbol{\beta}}$. In practice, however, theoretical expression of the covariance matrix is quite cumbersome, and bootstrap techniques are used. As in model D1, data must be resampled depending on the study design and the presence of clusters.

Estimation of baseline function $h_0(\cdot)$

Once the regression parameters $\boldsymbol{\beta}$ have been estimated, $h_0(\cdot)$ must in turn be estimated in order to estimate the covariate-specific ROC curve (4). For a fixed $t \in (0, 1)$, estimation of $h_0(t)$ can be made (see expressions in (5)) based on binary placement value indicators $\hat{B}_{jt} = I[PV_j \leq t], j = 1, \ldots, n_D$ by solving the following estimation equation:

$$\sum_{j=1}^{n_D} \omega_g(h_0(t) + \boldsymbol{x}_{Dj}'\hat{\boldsymbol{\beta}})\{\hat{B}_{jt} - g(h_0(t) + \boldsymbol{x}_{Dj}'\hat{\boldsymbol{\beta}})\} = 0.$$

It should be noted that the consistency of $\hat{h}_0(t)$ is obtained by ensuring that $t \in [a, b]$, with $a, b \in (0, 1)$, such that $P[PV_1 > a] > 0$ and $P[PV_1 < b] > 0$ (Cai, 2004).

3. Simulation studies

This section reports the results of a simulation study comparing the ROC approaches reviewed in Section 2. Different scenarios and different implementations of the methods outlined in the above section have been taken into account in this study. The aim was to study the performance of the different methods, by evaluating their robustness in the face of deviations from the theoretical model.

Data were simulated from two scenarios, namely,

- Scenario I

$$Y_D = 2 + 4X + \varepsilon_D \tag{8a}$$

and

$$Y_{\bar{D}} = 1.5 + 3X + \varepsilon_{\bar{D}}. \tag{8b}$$

- Scenario II

$$Y_D = 5 + 3X^2 - 25\,(X - 0.2)_+^3 + 250\,(X - 0.65)_+^3 + \varepsilon_D \tag{9a}$$

and

$$Y_{\bar{D}} = 4 + 0.6X + \varepsilon_{\bar{D}}. \tag{9b}$$

In both scenarios, X was generated from $U(0, 1)$. Bearing in mind the distribution of errors ε_D and $\varepsilon_{\bar{D}}$, the following situations were considered: (a) ε_D and $\varepsilon_{\bar{D}}$ displaying normal distributions; (b) ε_D and $\varepsilon_{\bar{D}}$ displaying extreme value distributions; (c) ε_D and $\varepsilon_{\bar{D}}$ displaying t-distributions; and, (d) ε_D and $\varepsilon_{\bar{D}}$ displaying logistic distributions. In all cases, the errors were generated with a mean of zero and a standard deviation of 1.5, for healthy and diseased populations.

The discrepancy between the estimator of the covariate-specific ROC curve and the true ROC curve was measured in terms of the empirical version of the global mean squared error (MSE):

$$\text{MSE} = \frac{1}{n_X}\sum_{l=0}^{n_X}\frac{1}{n_T}\sum_{r=0}^{n_T}(\widehat{\text{ROC}}_{X=x_l}(t_r) - ROC_{X=x_l}(t_r))^2,$$

with $x_l = \frac{l}{n_X}$, $l = 0, \ldots, n_X$, $t_r = \frac{r}{n_T}$, $r = 0, \ldots, n_T$, and $n_X = n_T = 50$.

With respect to direct methodology, in all the results shown below, the probit function, namely, $g^{-1} = \Phi^{-1}$, was taken as the link function. Similarly, two possible estimators for the conditional survival function in the healthy population $S_{\bar{D}X}$ were studied. In both cases, a parametric location-scale model for the healthy population, $Y_{\bar{D}} = \mu(X) + \sigma\varepsilon$, was assumed, such that $S_{\bar{D}X}(y) = S_0(\frac{y-\mu(X)}{\sigma})$, where S_0 is the survival function of error ε. In line with the assumptions made about the distribution of ε, estimators will be referred to as: (a) 'normal', where Gaussian error is assumed, i.e., $S_0(y) = 1 - \Phi(y)$; and, (b) 'semiparametric', where no assumption is made about distribution. In such a case, the survival function S_0 was empirically estimated on the basis of standardised residuals.

Finally, for the D1 approach, it was assumed that $h_0(t) = \alpha_0 + \alpha_1\Phi^{-1}(t)$, with (α_0, α_1) unknown. With respect to the set of FPFs, $n_T = 50$ and equally spaced values were considered.

An important point when working with parametric regression models is the modelling of covariate effects. In many situations, the assumption of linearity is not verified, thus making it necessary for flexibility to be incorporated into regression models. As pointed out above, a possible option, in all cases within a parametric framework, is the use of regression splines, such as B-splines or natural splines. Accordingly, to evaluate the impact of misspecification of covariate's effect on the ROC curve, in Scenario II of this study two models were fitted to simulated data, i.e., a linear model and a spline model, with cubic B-splines as bases, and knots at 0.2 and 0.65.

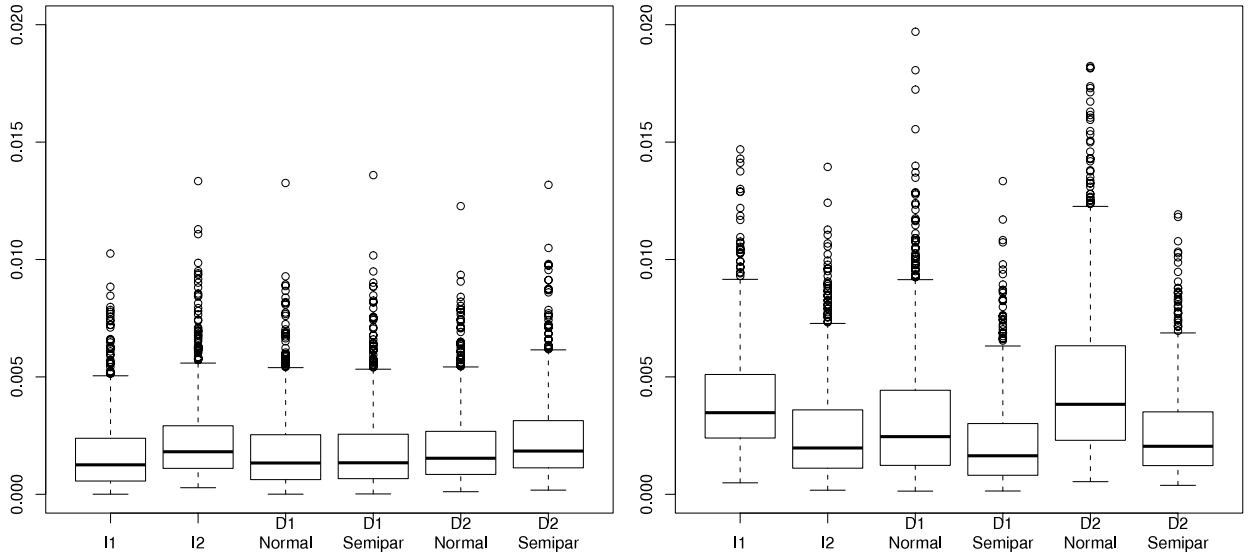
The results of the simulation study pertaining to Gaussian and extreme value distributed errors are shown below; the results for the remaining distributions can be found in the Appendix. In every case, the same sample size was considered for both diseased and healthy subjects, with $n_D = n_{\bar{D}} = 50$, 200, 500, and the results shown are based on 1000 simulated data sets.

3.1. Normal errors

Under the assumption of Gaussian errors in (8) and (9), $\varepsilon_D, \varepsilon_{\bar{D}} \sim N(0, 1.5^2)$, the true covariate-specific ROC curve for each of the scenarios considered, is:

I. $ROC_X(t) = \Phi(\frac{0.5+X}{1.5} + \Phi^{-1}(t))$.

II. $ROC_X(t) = \Phi(\frac{1-0.6X+3X^2-25(X-0.2)_+^3+250(X-0.65)_+^3}{1.5} + \Phi^{-1}(t))$.



(a) Normal errors.

(b) Extreme value distributed errors.

Fig. 1. Boxplots of the estimated MSE for scenario I and $n_D = n_{\bar{D}} = 200$ based on 1000 estimates.

Table 1(a) shows the mean and standard deviation of the MSE obtained in 1000 simulations, for all the models considered. In every case, the same sample size was considered for both diseased and healthy subjects, with $n_D = n_{\bar{D}} = 50, 200, 500$. As will be seen, model I1 performed best in all cases, though the performance of the remaining models can be deemed satisfactory. One exception to this occurred in the linear fit for Scenario II. As was to be expected, and for all models indiscriminately, incorrect modelling of the covariate effect had a great impact on the MSE. If one focuses on the direct methodology, it is interesting to note the effect of choice of estimator of the conditional survival function on the healthy population, $S_{\bar{D}X}$. As can be seen, the choice of the semiparametric estimator yields results, in terms of MSE, similar to the case of the normal estimator, with these differences being minimal in model D1. In our opinion, this suggests the use in practice of the semipar version of models D1 and D2, in that fewer assumptions are made than in the normal case. Finally, the MSE decreases when sample size increases. All the above is clearly observable in Figs. 1(a) and 2, in which the boxplot of the 1000 estimated MSEs and the average of the estimated AUCs, along with 2.5 and 97.5 simulation quantiles, for Scenario I and sample size 200, are shown.

3.2. Extreme value errors

Assuming that the errors in (8) and (9) follow an extreme value distribution, with a zero mean and standard deviation of 1.5, the true covariate-specific ROC curve for each of the scenarios considered, is:

I. $ROC_X(t) = 1 - (1-t)\exp(\frac{0.5+X}{1.5}\sqrt{6/\pi})$.

II. $ROC_X(t) = 1 - (1-t)\exp(\frac{1-0.6X+3X^2-25(X-0.2)_+^3+250(X-0.65)_+^3}{1.5}\sqrt{6/\pi})$.

The results in terms of the MSE, obtained in 1000 simulations, are shown in Table 1. As was to be expected, the MSE decreased when sample size increased. Fig. 1(b) depicts the boxplot of the 1000 estimated MSEs, and Fig. 3 shows the average of the estimated AUC, along with 2.5 and 97.5 simulation quantiles, for Scenario I and sample size 200. As can be seen in Fig. 3, while in terms of MSE, essentially for small sample sizes, differences between the respective methods were not that evident, neither model I1 nor direct modelling based on normal estimation of the conditional survival function performed satisfactorily, inasmuch as they furnished biased estimations. Direct modelling with semiparametric estimation of the survival function $S_{\bar{D}X}$ warrants special attention. Interestingly, the performance of the estimators was generally acceptable, though neither the link function nor, in the case of model D1, the baseline function of the false positives was well specified. These results suggest that semipar versions of models D1 and D2, are robust in the face of misspecifications of the ROC regression model, with the exception of incorrect modelling of the covariate effect. With regard to this last point, as in the case of Gaussian errors, incorrect modelling of the covariate effect had a great impact on the MSE for all models. Finally, as was to be expected, the model I2 performed well. This model is the only one in respect of which no distribution assumption whatsoever, implicit or express, was made.

3.3. Other errors

It should be pointed out that, as in the case of extreme value distributed errors, the robustness of direct modelling (with semiparametric estimation of the survival function in healthy subjects) vis-à-vis misspecifications of the ROC regression

Table 1
Average (standard deviation) of estimated mean squared error (MSE) ($\times$ 1000) over the 1000 simulated data sets.

Scenario		Sample size $n_D = n_{\tilde{D}}$	I1	I2	D1 normal	D1 semipar	D2 normal	D2 semipar
(a) Normal errors								
I		50	6.995 (6.649)	9.160 (7.224)	7.509 (7.030)	7.422 (6.966)	8.202 (6.939)	9.648 (7.331)
		200	1.690 (1.527)	2.291 (1.713)	1.817 (1.629)	1.855 (1.674)	1.998 (1.579)	2.347 (1.702)
		500	0.672 (0.647)	0.890 (0.685)	0.710 (0.671)	0.722 (0.673)	0.791 (0.673)	0.914 (0.696)
II	Linear	50	26.118 (6.039)	28.529 (6.866)	26.803 (6.533)	26.686 (6.485)	27.468 (6.392)	28.768 (6.700)
		200	21.342 (1.554)	21.947 (1.759)	21.361 (1.537)	21.368 (1.541)	21.442 (1.528)	21.721 (1.585)
		500	20.346 (0.692)	20.592 (0.765)	20.304 (0.644)	20.313 (0.651)	20.248 (0.644)	20.365 (0.663)
	Spline	50	15.168 (10.049)	18.232 (10.965)	18.164 (11.323)	18.016 (11.319)	18.375 (10.944)	19.852 (11.231)
		200	3.471 (2.116)	4.103 (2.213)	3.859 (2.383)	3.876 (2.394)	4.019 (2.279)	4.312 (2.326)
		500	1.382 (0.833)	1.619 (0.864)	1.526 (0.892)	1.535 (0.900)	1.597 (0.869)	1.714 (0.892)
(b) Extreme value distributed errors								
I		50	10.023 (7.814)	10.202 (7.824)	9.516 (8.856)	8.023 (7.190)	11.514 (9.580)	10.001 (7.294)
		200	3.982 (2.239)	2.637 (2.043)	3.240 (2.744)	2.169 (1.842)	4.795 (3.375)	2.597 (1.863)
		500	2.587 (1.015)	1.020 (0.781)	1.849 (1.426)	0.924 (0.723)	3.239 (1.862)	1.081 (0.689)
II	Linear	50	32.942 (6.782)	32.884 (7.338)	33.476 (7.308)	32.755 (6.356)	34.477 (7.545)	33.491 (6.856)
		200	27.842 (1.941)	25.952 (2.078)	28.053 (2.152)	27.555 (1.729)	28.566 (2.582)	26.449 (1.967)
		500	26.756 (0.929)	24.419 (0.954)	26.897 (1.088)	26.452 (0.702)	27.196 (1.391)	24.975 (0.841)
	Spline	50	18.194 (10.161)	19.718 (10.409)	20.111 (11.387)	18.996 (10.780)	21.308 (11.071)	20.407 (10.595)
		200	6.053 (2.613)	5.001 (2.471)	6.280 (2.949)	5.540 (2.515)	7.256 (3.218)	5.204 (2.579)
		500	3.722 (1.201)	2.128 (0.946)	3.861 (1.429)	3.244 (1.035)	4.634 (1.653)	2.559 (1.091)

model fitted, was observed in simulations performed under other distribution assumptions. In the case of logistic errors, however, the performance of models I1, D1 Normal and D2 Normal can be deemed satisfactory. In contrast, in the case of errors displaying a Student's t distribution, the differences among the various models were, if anything, more pronounced. A detailed summary of all these results can be found in the Appendix.

4. Application to a CAD system

Breast cancer continues to be one of the most common cancers, and survival rates critically depend on detection in the initial stages (Jemal et al., 2008). Several studies have demonstrated the benefits and potential of using CAD systems to help specialists in their clinical interpretation of mammograms (Nishikawa, 2007). As an intermediate step in diagnosis, a CAD system generates a set of suspicious regions extracted from the image, to be definitively classified as lesions (e.g masses) or false detections (FDs).

According to the American College of Radiology (2003), masses can be defined as three-dimensional structures demonstrating convex outward borders, usually evident on two orthogonal views. In terms of radiographic image findings, those structures can be characterized by a gradual variation of the pixel value from the border to the center. This variation is a consequence of the projection of a 3D structure (a mass) onto a 2D image plane. To model this behavior we used the iris filter, which is capable of enhancing those type of structures.

When this type of filter is applied to real masses, this brings a new factor into play that is linked to the tissue surrounding the lesion. If the tissue is predominantly fatty, the mass is easily detectable, owing to the high grey level of the mass with respect to the surrounding tissue. If the tissue is dense, however, its average grey level will also be high, thereby reducing the contrast between any possible mass and such tissue. All this serves to hinder detection of any possible mass.

The methodologies reviewed in this paper were applied to a CAD system dedicated to the detection of breast masses. The objective has been to assess the effect on the accuracy of the iris filter when discriminating between real masses and

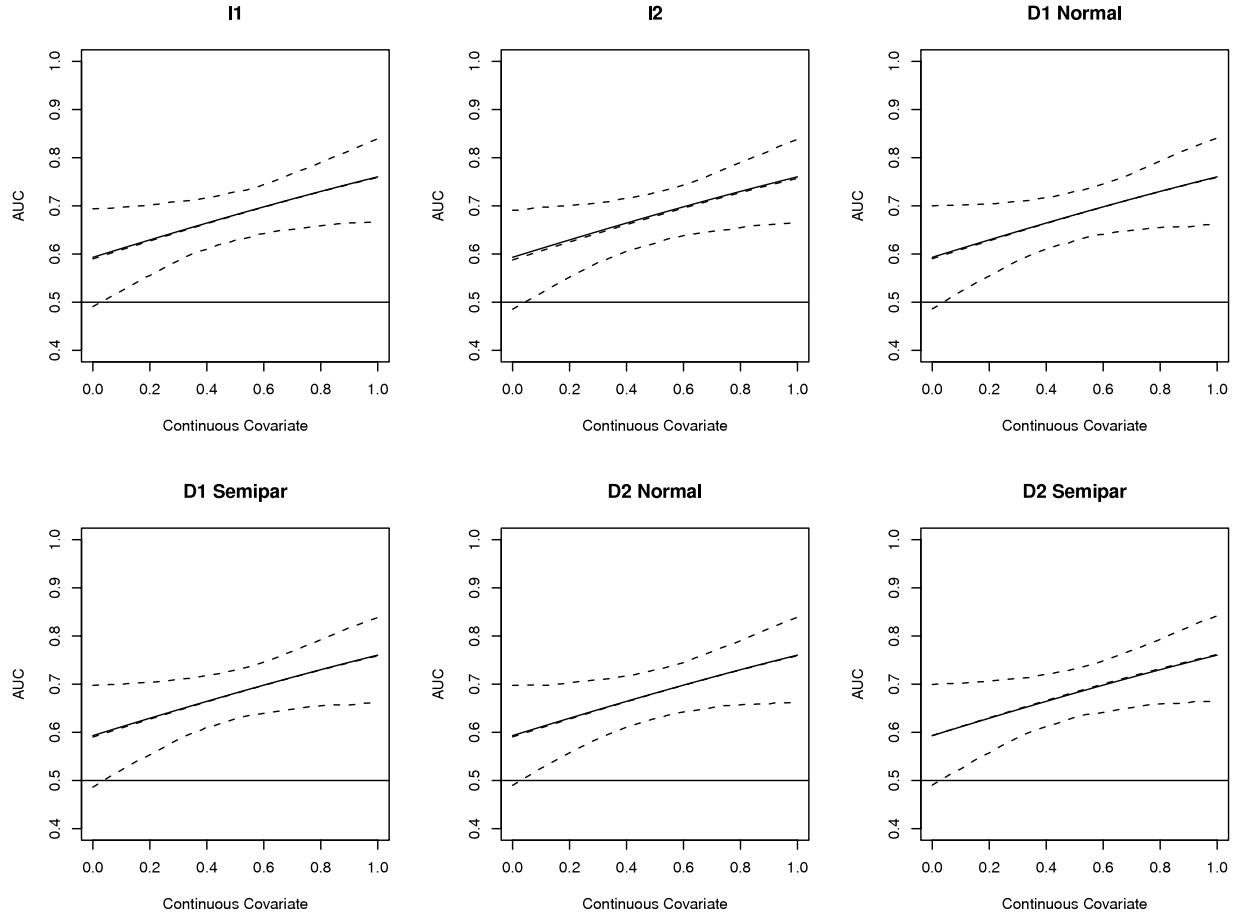


Fig. 2. True area under the ROC curve (AUC) (solid line) versus average of simulated AUCs (dashed line), along with 2.5 and 97.5 simulation quantiles for scenario I (normal errors) and $n_D = n_{\bar{D}} = 200$ based on 1000 estimates.

false detections of (a) the average grey level of the pixels forming the suspicious region, (b) the eccentricity of the suspicious region, and (c) the breast tissue type.

4.1. Data set

The CAD scheme (see Varela et al. (2007) for a detailed description) was applied to a database containing 580 mammograms where a total of 190 images were classified as abnormal (lesion present), and the remaining 390 as normal (no lesion present). From the 580 original mammograms, a total of 2796 regions suspicious of being a malignant mass were detected by the computer in a first step. Of these, only 384 corresponded to true masses, and the remainder, a total of 2412, corresponded to false detections.

4.2. Data analysis

In the first stage of the analysis, the effect of covariate eccentricity on the iris filter's (Y) capacity to discriminate between malignant masses (D) and false detections ($\bar{D}$) was assessed. In all such analyses, breast tissue type (dense or fatty) was also included. With the aim of comparing the performance, in practice, of the different methodologies outlined in Section 2, all of these were applied during this stage of the study.

As a first step, separately exploratory analyses of the data set were performed for masses and FDs respectively, using smooth additive models (AM) (Wood, 2006). Fig. 4 shows the effect of eccentricity on the iris filter, according to breast tissue type. As seen, in the case of fatty tissue, the iris filter value was higher for masses than for FDs. Moreover, while the effect of eccentricity on masses was observed to be nil, this was not so in the case of FDs. As eccentricity increases (round shape), the output of the filter rises, as expected. With dense tissue, in both cases (masses and FDs) filter output was observed to be higher than before.

The results of the ROC regression analysis are shown below. Previous exploratory analyses seemed to indicate the existence of interaction between eccentricity and tissue type for FDs. Similarly, the effect of eccentricity on the iris filter

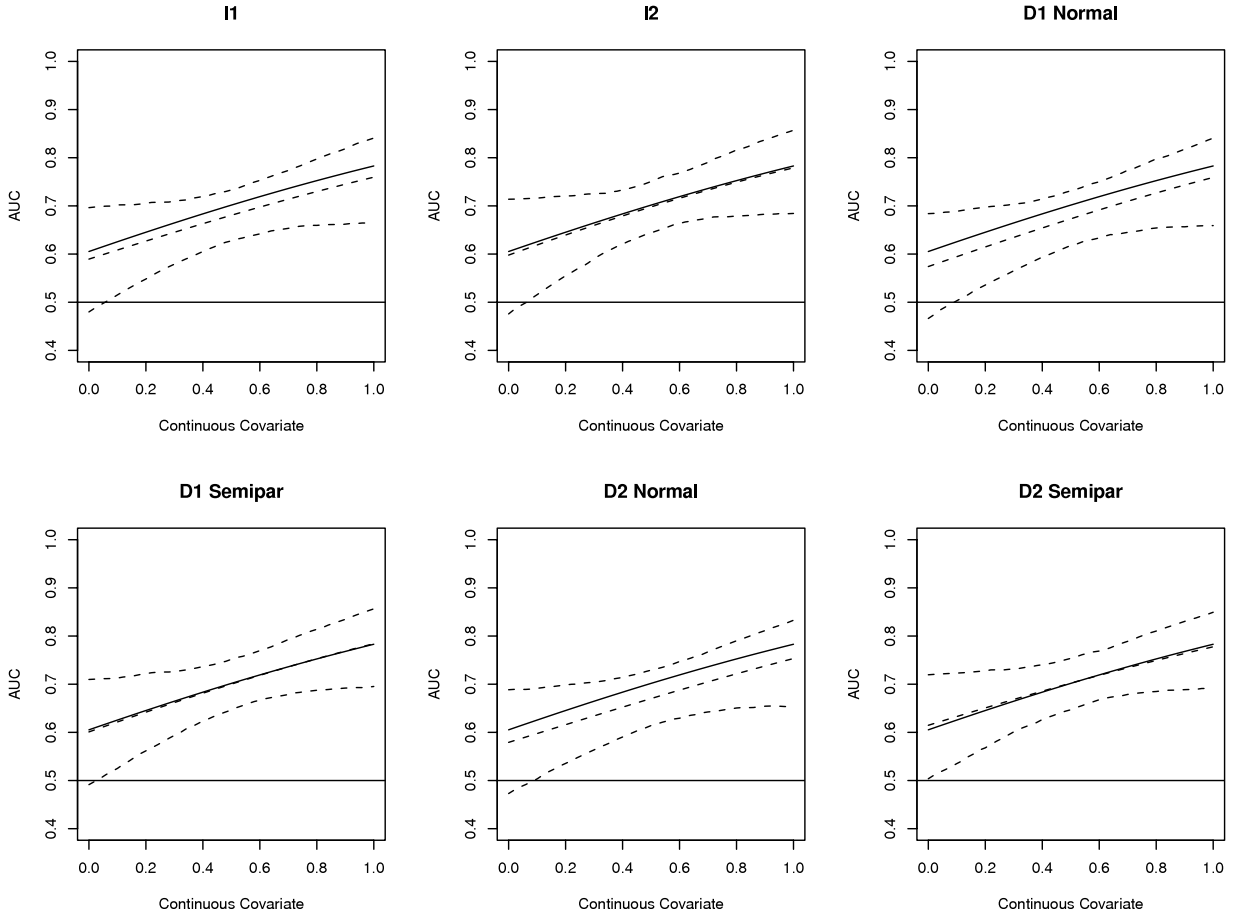


Fig. 3. True area under the ROC curve (AUC) (solid line) versus average of simulated AUCs (dashed line), along with 2.5 and 97.5 simulation quantiles for scenario I (extreme value errors) and $n_D = n_{\bar{D}} = 200$ based on 1000 estimates.

proved to be approximately linear for both dense and fatty tissue (see Fig. 4). Accordingly, the following models in the case of masses and FDs were considered for induced methodology

$$E\left[Y_D|\mathrm{ECC}, \mathrm{TIS}\right] = \beta_{0D} + \beta_{1D}\mathrm{ECC} + \beta_{2D}\mathrm{TIS}$$

and

$$E\left[Y_{\bar{D}}|\mathrm{ECC}, \mathrm{TIS}\right] = \beta_{0\bar{D}} + \beta_{1\bar{D}}\mathrm{ECC} + \beta_{2\bar{D}}\mathrm{TIS} + \beta_{3\bar{D}}\mathrm{ECC} \times \mathrm{TIS}, \tag{10}$$

where Y_D and $Y_{\bar{D}}$ denote filter output on masses and FDs respectively, ECC denotes covariate eccentricity, and TIS is a dummy variable, taking a value of 1 in the case of fatty tissue and 0 in the case of dense tissue. An analysis of the residuals of the previous regression models reveals that both the normality and homoscedasticity assumptions are then plausible for our data.

For direct methodology, the model fitted for estimating the conditional survival function in the FD population was the same as in the previous case (see (10)). As regards the ROC curve, this was modelled as

$$ROC_{(\mathrm{ECC,TIS})} = \Phi\left(h_0(t) + \beta_1\mathrm{ECC} + \beta_2\mathrm{TIS} + \beta_3\mathrm{ECC} \times \mathrm{TIS}\right).$$

The estimated ROC regression parameters, along with their 95% confidence intervals, are shown in Table 2. For model I1, confidence intervals were calculated according to the procedure indicated in Section 2.1. For the remaining models, bootstrap techniques, with 500 resamples, were used to compute the standard errors of the ROC parameter estimates. The confidence intervals were then calculated as per the normal approach. As can be seen, all the methods yielded similar results. The iris filter's discriminatory capacity decreased as eccentricity increased, with this decline being far more pronounced in the case of dense tissue. This pattern is more clearly observable in Fig. 5, where the covariate-specific AUC for the model I1 is shown (the remaining models furnished similar results). For this case, the confidence intervals were calculated based on Faraggi's results (Faraggi, 2003). One aspect deserving comment are the values of the AUC for the dense tissue which are below 0.5. From a practical viewpoint, an AUC below 0.5 indicates that the usual classification rule ($Y \geq c$ indicates disease, $Y < c$ indicates health) is not appropriate and, therefore, must change ($Y < c$ indicates disease, $Y \geq c$ indicates health).

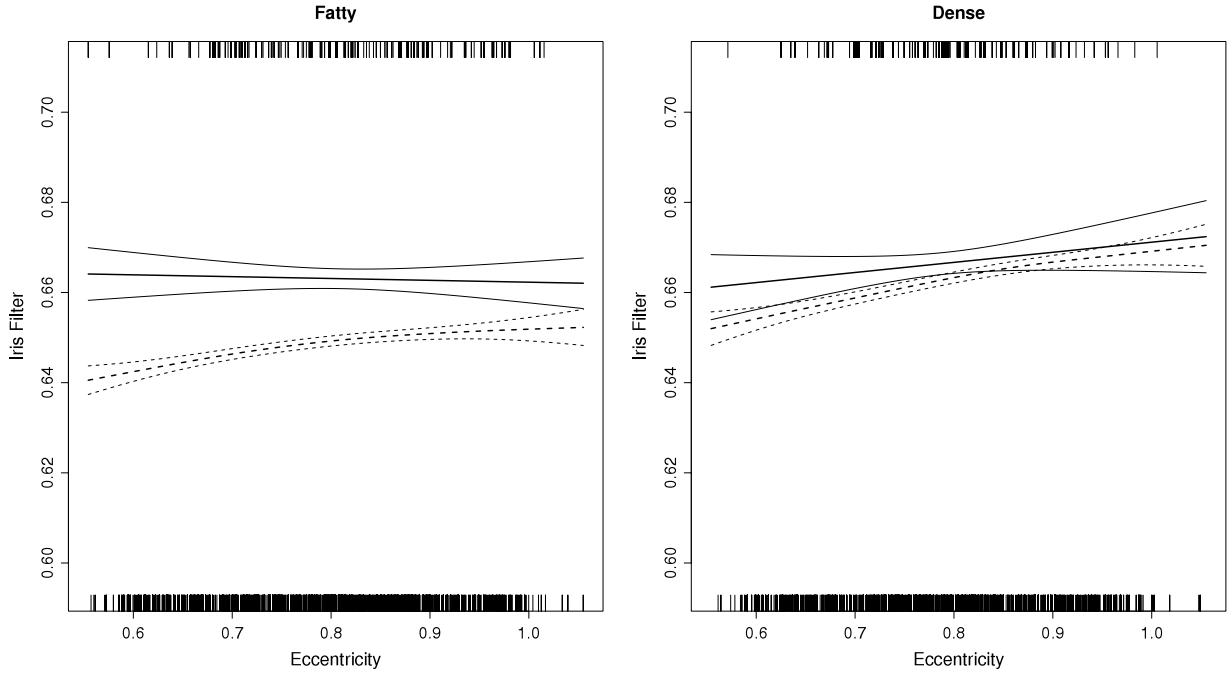


Fig. 4. Non-parametric estimates of iris filter, according to eccentricity, along with 95% pointwise confidence bands, for fatty and dense tissue. Solid line: masses. Dashed line: false detections (FDs).

Table 2
Estimated ROC regression parameters for the iris filter and eccentricity. In model I2, the ROC curve is expressed by reference to the survival function. In this case the estimated parameters of the ROC curve thus display the opposite sign.

Methodology	Model	Covariate	Estimate	95% confidence interval	*p*-value
Induced	I1	ECC	−2.23	(−3.462, −0.995)	0.0004
		TIS	−0.09	(−0.728, 0.551)	0.7868
		ECC × TIS	0.98	(0.221, 1.742)	0.0114
	I2	ECC	2.23	(0.100, 3.456)	0.0004
		TIS	0.09	(−0.609, 0.785)	0.8040
		ECC × TIS	−0.98	(−1.811, −0.152)	0.0203
Direct	D1 normal	ECC	−2.08	(−3.315, −0.852)	0.0009
		TIS	−0.14	(−0.774, 0.494)	0.6648
		ECC × TIS	1.13	(0.368, 1.895)	0.0037
	D1 semipar	ECC	−2.10	(−3.356, −0.846)	0.0010
		TIS	−0.15	(−0.835, 0.538)	0.6718
		ECC × TIS	1.14	(0.292, 1.978)	0.0083
	D2 normal	ECC	−2.29	(−3.554, −1.017)	0.0004
		TIS	−0.26	(−0.930, 0.416)	0.4549
		ECC × TIS	1.24	(0.430, 2.057)	0.0027
	D2 semipar	ECC	−2.29	(−3.536, −1.035)	0.0003
		TIS	−0.26	(−0.973, 0.460)	0.4836
		ECC × TIS	1.24	(0.387, 2.101)	0.0045

Thus, from the eccentricity value of 0.9 onwards the classification rule should be that high iris filter values are indicative of false detections (FD) instead of true masses.

In the second stage of the analysis, the effect of average grey level value on the iris filter's discriminatory capacity was assessed. As in the previous case, tissue type was included in all the analyses. Fig. 6 depicts the effect of average grey level value on filter output, according to breast tissue type. For masses, in both fatty and dense tissues, values rose to a peak approximately midway through the interval and fell thereafter. As a feature that measures the gradual variation in the region's grey level value, filter output tends to rise to a maximum in these intermediate areas, since it is here that such variation could register its most extreme values. For FDs, the pattern was more homogeneous, owing to the fact that, ideally, grey level values display no gradual variation and are instead homogeneously distributed.

From Fig. 6, it is evident that the effect of average grey level values on the iris filter is clearly non-linear. It is thus to be expected that the above pattern would also be present in the ROC curve. Since direct methodology makes the modelling of non-linear effects a complicated task (see Section 4.3), induced methodology was chosen as the modelling strategy for this

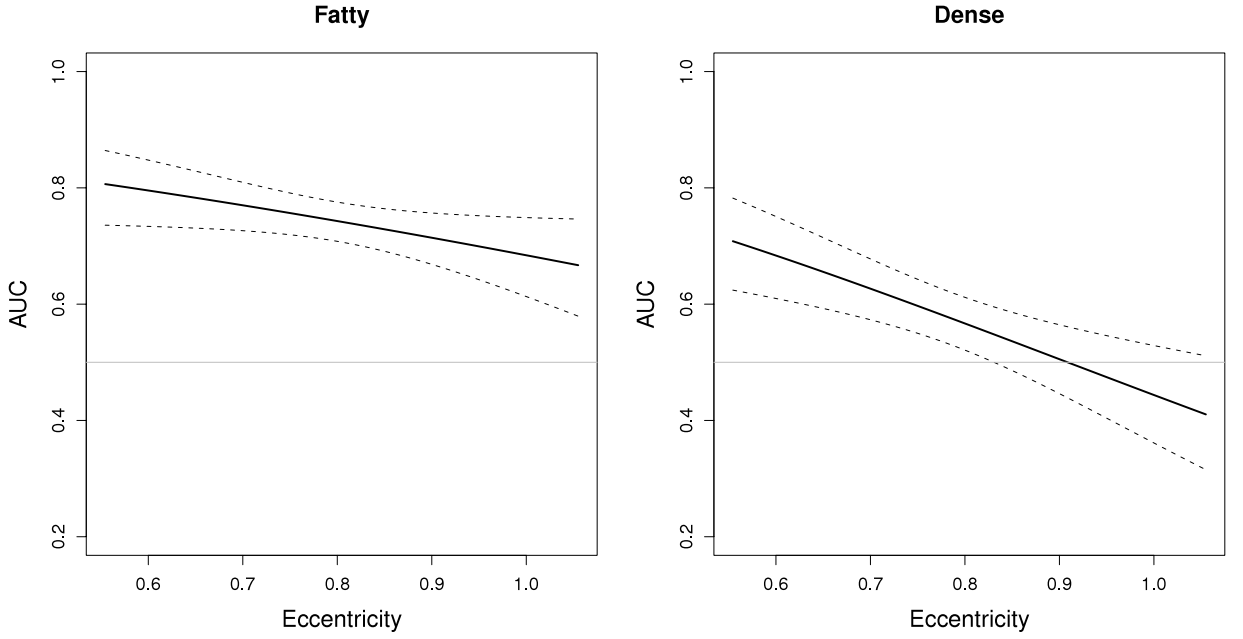


Fig. 5. Estimated AUC of iris filter, according to eccentricity, along with 95% pointwise confidence bands, for fatty and dense tissue.

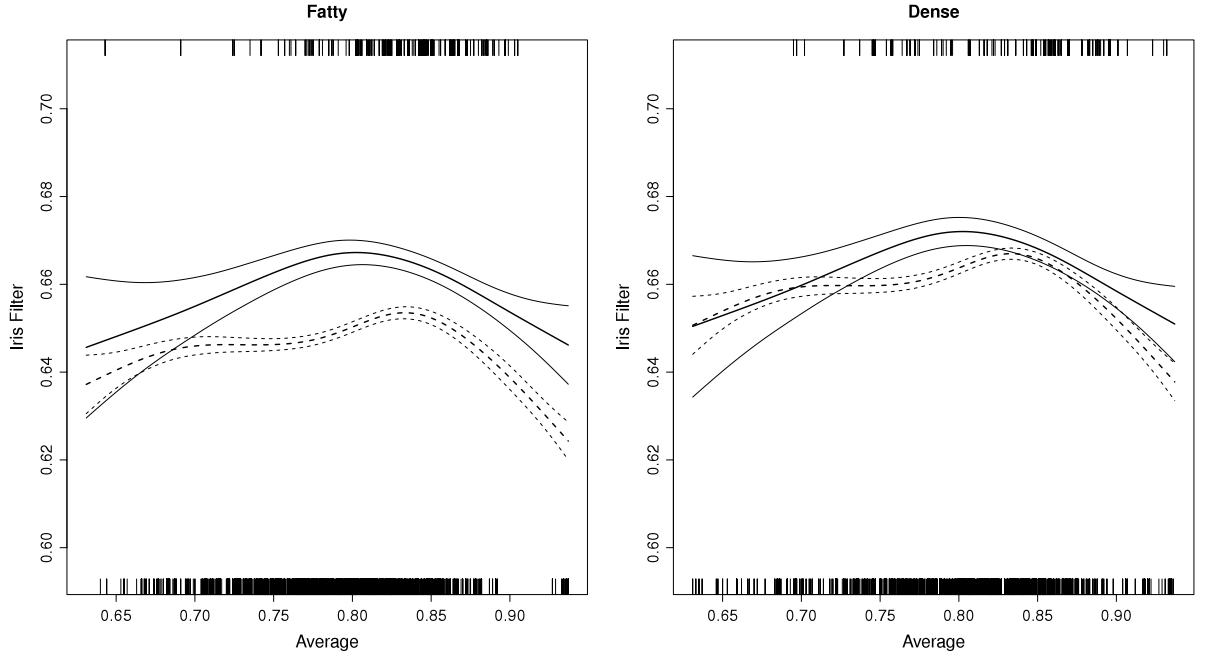


Fig. 6. Non-parametric estimates of iris filter, according to average, along with 95% pointwise confidence bands, for fatty and dense tissue. Solid line: masses. Dashed line: false detections (FDs).

analysis. For masses, the following model was considered:

$$E\left[Y_D|\text{AVG, TIS}\right] = \beta_{0D} + \boldsymbol{\beta}_{1D}B\left(\text{AVG}\right) + \beta_{2D}\text{TIS},$$

where AVG denotes the covariate average, TIS is a dummy variable, taking the value of 1 in the case of fatty tissue and 0 in the case of dense tissue, and B(AVG) is a basis for ordinary cubic polynomial regression.

In the case of FDs, the mean function was modelled as

$$E\left[Y_{\bar{D}}|\text{AVG, TIS}\right] = \beta_{0\bar{D}} + \boldsymbol{\beta}_{1\bar{D}}B\left(\text{AVG}\right) + \beta_{2\bar{D}}\text{TIS},$$

where B(AVG) is a cubic B-spline basis with knots at 0.8 and 0.83.

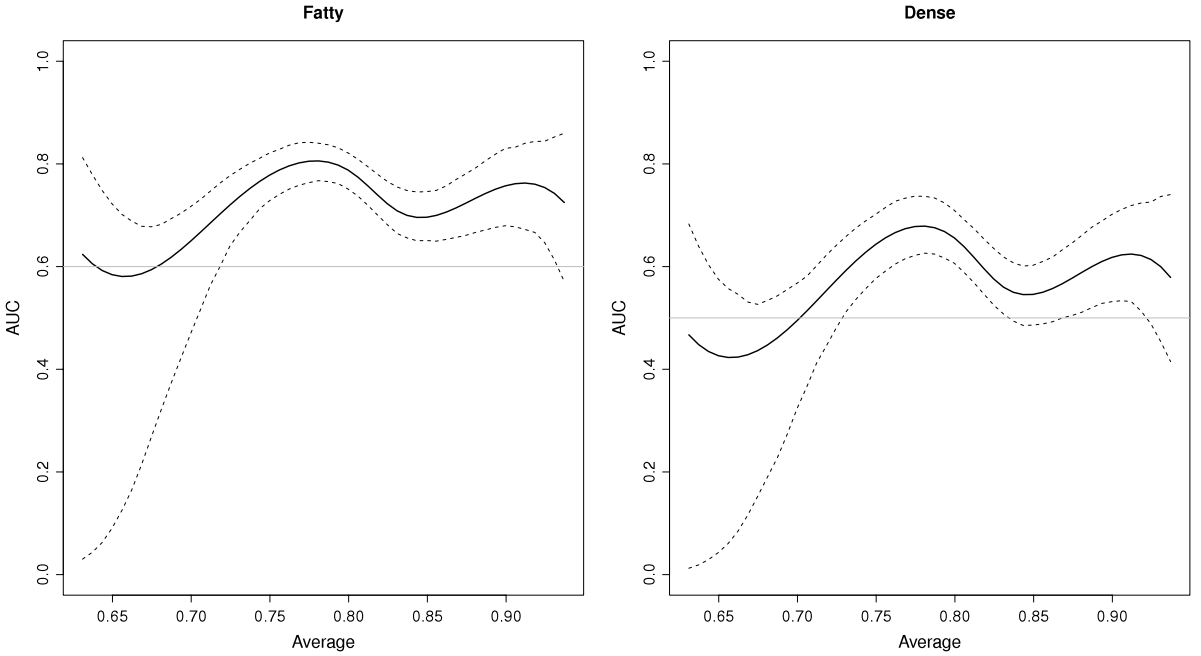


Fig. 7. Estimated AUC of iris filter, according to average, along with 95% pointwise bootstrap confidence bands, for fatty and dense tissue.

Finally, to estimate the ROC curve, model I2 was applied. Fig. 7 depicts the AUC estimated by reference to average grey level values, in dense and fatty tissue respectively. As can be seen, the effect of grey level on the iris filter's discriminatory capacity was similar in both cases, albeit with systematically higher values in the case of fatty tissue. For both types of tissue, however, the greatest discriminatory capacity was observed at average grey level values of close on 0.80. As stated above, these intermediate values are those where there can assumed to be a range of values specific to the characteristics of masses.

4.3. Model checking

For induced methodology, familiar model checking techniques can be used to validate the regression model assumed for the marker in both healthy, $Y_{\bar{D}}$, and diseased, Y_D, populations. In this application, the adequacy of the regression models was evaluated by an analysis of the residuals. In the case of direct methodology, however, this is a much more complicated task and very few contributions have been made to the literature so far (see e.g. Cai and Zheng (2007)). In fact, no procedures have been yet proposed, for instance, to assess the assumed functional form for the effect of continuous covariates on the ROC curve. In this work, induced methodology was used as an 'informal' diagnostic tool. Based on the estimated 'induced' effect of the continuous covariate on the ROC, we have modelled the effect of the covariate 'directly' on the ROC curve.

5. Discussion

This study sought to review different approaches to the inclusion of covariates in the context of ROC analysis and its application to the field of radiodiagnosis, specifically in the area of automatic detection of tumour masses in digital mammograms.

Specifically, two major paradigms were reviewed within the context of ROC regression: (a) induced methodology and (b) direct methodology. For induced methodology, we considered two possible implementations that differed in their assumptions about the distribution of errors in regression models for diseased and healthy subjects. In the case of Gaussian errors, the theoretical distribution of ROC regression parameters is presented in this study. Where no assumption was made about errors, we extended Pepe's study (Pepe, 1998) to allow for different distributions in diseased and healthy subjects.

The simulation study in Section 3 shows that both methodologies perform similarly assuming normality in the test result for diseased and healthy subjects. In the case of direct methodology, the study reveals no important effect of the estimator of the conditional survival function in healthy subjects, $S_{\bar{D}\boldsymbol{X}}$. As expected, incorrect modelling of covariate effects has an important effect on the final estimates. Under general conditions (i.e., non-normality scenarios), neither the I1 model nor, in direct modelling, implementations based on normal estimation of $S_{\bar{D}\boldsymbol{X}}$ yield satisfactory results. It is important to point out that the results of the simulation study reported in this study, as well as other analyses performed (not shown here), suggest that direct modelling (with semiparametric estimation of the survival function in healthy subjects) is robust vis-à-vis ROC regression model misspecifications of, say, the link function, or, in the case of model D1, the baseline function of false positives.

The advantage of induced modelling is that familiar statistical techniques can be used, since it is based on modelling the marker in diseased and healthy subjects separately. Nevertheless, direct modelling offers two major advantages in that it enables (a) the interaction between covariates and the false positive fraction to be easily incorporated and (b) different markers to be compared by simply incorporating the type of test as a categorical covariate in the ROC regression model. With respect to point (a) above, in this study we focused on induced methodology with homoscedastic regression models. Thought could, however, be given allowing for heteroscedastic models (see, for example, Zheng and Heagerty (2004)). This new configuration implicitly incorporates a possible interaction between the false positive fraction and covariates in the induced ROC curve.

Insofar as its application to the development of CAD systems is concerned, the results of this study would allow for the performance of classification algorithms to be assessed when they are applied to suspicious regions. This would enable existing interpretation of the performance of such algorithms to be improved, and the question of whether or not interactions exist to be analysed. Accordingly, it is possible to propose design alternatives that would allow for more complex and useful algorithms to be developed.

Finally, as shown in the second analysis, a crucial point when applying ROC regression methodologies is how the non-linear effect of continuous covariates on the ROC curve can be modelled parametrically. From an applied standpoint, principally in the case of direct modelling, this may make ROC analysis a complicated task. In this study we used B-splines, although this approach is not altogether satisfactory, due to the need to choose the number and location of knots. Although in the non-parametric framework some attempt has been made for induced methodology (González-Manteiga et al., 2010; Rodríguez-Álvarez et al., 2010), much work is still needed. Thus, an interesting field for future research would be to extend the methodologies described in this study to the non-parametric context, by the use, for instance, of penalised splines (Eilers and Marx, 1996) or the Bayesian versions of them (Lang and Brezger, 2004).

All the statistical analyses have been conducted with the ROC Regression package. This package can be obtained by contacting the first author at mariajose.rodriguez.alvarez@usc.es.

Acknowledgements

The authors would like to express their gratitude for the support received in the form of the Spanish MEC Grant MTM2008-01603 and the Galician Regional Authority (Xunta de Galicia) project INCITE08PXIB208113PR. We are also grateful to the two peer referees for their valuable comments and suggestions, which served to make a substantial improvement to this paper.

Appendix. Supplementary data

Supplementary material related to this article can be found online at doi:10.1016/j.csda.2010.07.018.

References

Alonzo, T.A., Pepe, M.S., 2002. Distribution-free ROC analysis using binary regression techniques. Biostatistics 3, 421–432.
American College of Radiology, 2003. ACR BI-RADS — mammography. In: Ultrasound & Magnetic Resonance Imaging, 4th ed. American College of Radiology, Resto, VA.
Cai, T., 2004. Semi-parametric ROC regression analysis with placement values. Biostatistics 5, 45–60.
Cai, T., Pepe, M.S., 2002. Semiparametric receiver operating characteristic analysis to evaluate biomarkers for disease. Journal of the American Statistical Association 97, 1099–1107.
Cai, T., Zheng, Y., 2007. Model checking for ROC regression analysis. Biometrics 63, 152–163.
Cheng, S.C., Wei, L.J., Ying, Z., 1995. Analysis of transformation models with censored data. Biometrika 82, 835–845.
de Boor, C.A., 2001. A Practical Guide to Splines, revised ed. Springer-Verlag, New York.
Efron, B., Tibshirani, R.J., 1993. An Introduction to the Bootstrap. Chapman and Hall/CRP Press, New York.
Eilers, P., Marx, B., 1996. Flexible smoothing with B-splines and penalties. Statistical Science 11, 89–121.
Faraggi, D., 2003. Adjusting receiver operating characteristic curves and related indices for covariates. The Statistician 52, 179–192.
González-Manteiga, W., Pardo Fernández, J.C., Van Keilegom, I., 2010. ROC curves in nonparametric location-scale regression models. Scandinavian Journal of Statistics. doi:10.1111/j.1467-9469.2010.00693.x.
Hanley, J.A., Hajian-Tilaki, K.O., 1997. Sampling variability of non-parametric estimates of the area under receiver operating characteristic curves: an update. Academic Radiology 4, 49–58.
Janes, H., Pepe, M.S., 2009. Accommodating covariates in ROC analysis. Stata Journal 9, 17–39.
Jemal, A., Siegel, R., Ward, E., Hao, Y., Xu, J., Murray, T., Thun, M.J., 2008. Cancer statistics. CA: A Cancer Journal for Clinicians 58, 71–96.
Kshirsagar, A.M., 1961. Some extensions of the multivariate generalization t distribution and the multivariate generalization of the distribution of the regression coefficients. Proceedings of the Cambridge Philosophical Society 57, 80–85.
Lang, S., Brezger, A., 2004. Bayesian P-splines. Journal of Computational and Graphical Statistics 13, 183–212.
López de Ullibarri, I., Cao, R., Cadarso-Suárez, C., Lado, M.J., 2008a. Nonparametric estimation of conditional ROC curves: application to discrimination tasks in computerized detection of early breast cancer. Computational Statistics & Data Analysis 52, 2623–2631.
López de Ullibarri, I., Rodríguez-Álvarez, M.X., Cadarso-Suárez, C., 2008b. ROC.regression: a new R software for ROC regression analysis. In: Paul Eilers (Ed.), Proceedings of the 23rd International Workshop on Statistical Modelling, pp. 321–324.
McCullagh, P., Nelder, J.A., 1989. Generalized Linear Models, second ed.. Chapman and Hall, London.
Metz, C.E., 1978. Basic principles of ROC analysis. Seminars in Nuclear Medicine 8, 183–298.
Nishikawa, R.M., 2007. Current status and future directions of computer-aided diagnosis in mammography. Computerized Medical Imaging and Graphics 31, 224–235.
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., 2000. An interpretation for the ROC curve and inference using GLM procedures. Biometrics 56, 352–359.

Pepe, M.S., 2003. The Statistical Evaluation of Medical Tests for Classification and Prediction. Oxford University Press, New York.
Pepe, M.S., Cai, T., 2004. The analysis of placement values for evaluating discriminatory measures. Biometrics 60, 528–535.
R Development Core Team, 2010. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0, URL: http://www.R-project.org.
Rodríguez-Álvarez, M.X., Roca-Pardiñas, J., Cadarso-Suárez, C., 2010. ROC curve and covariates: extending induced methodology to the non-parametric framework. Statistics and Computing. doi:10.1007/s11222-010-9184-1.
Tosteson, A.N., Begg, C.B., 1988. A general regression methodology for ROC curve estimation. Medical Decision Making 8, 204–215.
Varela, C., Tahoces, P.G., Méndez, A.J., Souto, M., Vidal, J.J., 2007. Detection of breast tumours in digitized mammograms. Computers in Biology and Medicine 37, 214–226.
Venables, W., 1975. Calculating of confidence intervals for non-centrality parameters. Journal of the Royal Statistical Society, Series B 37, 406–412.
Wood, S.N., 2006. Generalized Additive Models. An Introduction with R. Chapman and Hall/CRP Press, New York.
Zheng, Y., Heagerty, P.J., 2004. Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. Biostatistics 4, 615–632.