Harvard University
Harvard University Biostatistics Working Paper Series

Model Checking for ROC Regression Analysis

Tianxi Cai[*]   Yingye Zheng[†]

[*]Harvard University, tcai@hsph.harvard.edu

[†]Fred Hutchinson Cancer Research Center, yzheng@fhcrc.org

Model Checking for ROC Regression Analysis

Tianxi Cai and Yingye Zheng

Abstract

The Receiver Operating Characteristic (ROC) curve is a prominent tool for characterizing the accuracy of continuous diagnostic test. To account for factors that might invluence the test accuracy, various ROC regression methods have been proposed. However, as in any regression analysis, when the assumed models do not fit the data well, these methods may render invalid and misleading results. To date practical model checking techniques suitable for validating existing ROC regression models are not yet available. In this paper, we develop cumulative residual based procedures to graphically and numerically assess the goodness-of-fit for some commonly used ROC regression models, and show how specific components of these models can be examined within this framework. We derive asymptotic null distributions for the residual process and discuss resampling procedures to approximate these distributions in practice. We illustrate our methods with a dataset from the Cystic Fibrosis registry.

Model Checking for ROC Regression Analysis

TIANXI CAI

Department of Biostatistics, Harvard University,

Boston, MA 02115

and YINGYE ZHENG

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,

Seattle, WA 98109

SUMMARY

The Receiver Operating Characteristic (ROC) curve is a prominent tool for characterizing the accuracy of a continuous diagnostic test. To account for factors that might influence the test accuracy, various ROC regression methods have been proposed. However, as in any regression analysis, when the assumed models do not fit the data well, these methods may render invalid and misleading results. To date practical model checking techniques suitable for validating existing ROC regression models are not yet available. In this paper, we develop cumulative residual based procedures to graphically and numerically assess the goodness-of-fit for some commonly used ROC regression models, and show how specific components of these models can be examined within this framework. We derive asymptotic null distributions for the residual processes and discuss resampling procedures to approximate these distributions in practice. We illustrate our methods with a dataset from the Cystic Fibrosis registry.

Key words : Cumulative Residual, Diagnostic Accuracy, Generalized Linear Model, Model Checking, ROC Regression

1. Introduction

The receiver operating characteristic (ROC) curve is a well accepted measure of accuracy for continuous diagnostic tests (Swets and Pickett, 1982). In recent years there has been much focus on developing ROC regression methodologies for evaluating factors that may influence the accuracy of a test. With more ROC regression procedures becoming available, it is tempting to apply these methods routinely in practice without considering the goodness of fit of a given model. However, as in any regression analysis, model checking is a crucial step in ROC regression since the validity of the statistical inference relies on the adequacy of model assumptions. To date model checking tools suitable for ROC regression are not well developed. In this paper we present model checking procedures that supplement some of the recently developed ROC regression methods.

Suppose a continuous test Y is used to predict a binary disease status D with $D = 1$ denoting diseased and $D = 0$ being disease-free. We use the convention that higher values of Y are more indicative of disease. The ROC curve is motivated as follows: if a threshold value c was used to classify subjects as diseased ($Y \geq c$) or disease-free ($Y < c$), then the true and false positive rates associated with this decision criterion are $\mathrm{TPR}(c) = \mathrm{P}(Y \geq c \mid D = 1) \equiv \mathrm{P}(Y_{\mathrm{D}} \geq c)$ and $\mathrm{FPR}(c) = \mathrm{P}(Y \geq c \mid D = 0) \equiv \mathrm{P}(Y_{\bar{\mathrm{D}}} \geq c)$, where Y_{D} and $Y_{\bar{\mathrm{D}}}$ denote the test results for diseased and non-diseased subjects, respectively. An ROC curve is a plot of $\{\mathrm{FPR}(c), \mathrm{TPR}(c)\}$ for $c \in (-\infty, \infty)$, or equivalently a function of the form $\mathrm{ROC}(u) = \mathrm{TPR}\{\mathrm{FPR}^{-1}(u)\}$ for $u \in (0, 1)$. It summarizes all of the compromises between increasing $\mathrm{TPR}(c)$ and simultaneously decreasing 1-$\mathrm{FPR}(c)$ as the threshold c is lowered. Thus the curve contains information necessary for choosing an appropriate threshold value c in any particular setting depending on the costs of false positives and false negatives. Another benefit of the ROC curve is that different diagnostic tests are placed on the same error rate scale therefore it allows the comparison of competing tests.

To evaluate possible covariate effects on the discriminatory capacity of a test, roughly three

2

different approaches to ROC regression have emerged from the literature: (1) modelling the distribution of Y_D and $Y_{\bar{D}}$ separately to induce regression models for the ROC curves (Tostenson and Begg, 1988; Toledano and Gatsonis, 1995; Zheng and Heagerty, 2004); (2) direct modelling of the ROC curves (Pepe, 1997, 2000; Alonzo and Pepe, 2002; Cai and Pepe, 2002; Izmirlian, 2003; Cai, 2004; Pepe and Cai, 2004); and 3) modelling summary accuracy indices such as the area under the ROC curves (AUC) or the partial area under the ROC curves (pAUC) (Thompson and Zuchini, 1989; Obuchowski, 1995; Dodd and Pepe, 2003a, 2003b; Cai and Dodd, 2004). Most existing ROC regression models, although closely related to generalized linear models (GLMs), are different from the ordinary GLM in both model specifications and inference procedures. For example, the semi-parametric location-scale model proposed by Heagerty and Pepe (1999) quantifies both the mean and variance as functions of covariates which results in more complex inference procedures compared to that for GLM. Within the direct modelling approaches, the parametric ROC-GLM (Pepe, 1997, 2000; Alonzo and Pepe, 2002; Pepe and Cai, 2004) and the semi-parametric ROC-GLM (Cai and Pepe, 2002; Cai, 2004) specify the *relationship* between the distributions of $Y_{\bar{D}}$ and Y_D using a GLM framework, but individual distributions do not necessarily follow GLMs. For these differences, existing model checking methods for GLMs do not apply directly and model validation for ROC regression models may be challenging.

There is a vast literature on model checking procedures for GLM. One popular approach is to use graphical residual analysis techniques (Cook and Weisberg, 1994, 1997). A plot of residuals against any coordinate, such as a covariate or the fitted value, may reveal functional misspecification of the covariates or lack of linearity in the mean. However, it is often difficult to ascertain the variability in a raw residual plot and thus one can only subjectively determine whether the observed residual pattern indicates anything beyond sampling variability. More analytic procedures based on certain aggregates of individual residuals have been proposed. One approach is

3

to partition the covariate space into distinct regions and form lack of fit tests based on grouped residuals (e.g., Tsiatis, 1980 and Landwehr, Pregibon and Shoemaker, 1984). However the partitions can be arbitrary and sometimes different partitions may lead to conflicting conclusions. To overcome these difficulties, Su and Wei (1991) proposed the use of cumulative residual processes as a more objective model checking tool and showed that these processes converge weakly to zero-mean Gaussian processes under the null of correct model specification. Kolmogorov-Smirnov supremum-type test statistics can be formed to objectively validate the assumed GLM. To more effectively examine specific components of a GLM, such as the link function or covariate functional forms, Lin, Wei and Ying (2002) considered a few variations of such processes by taking partial sums of residuals over different coordinates. Their approach is appealing as it facilitates the evaluation of specific model components both graphically and numerically.

The primary goal of this paper is to develop objective model checking techniques for some commonly used ROC regression models whose inference procedures are more complex than ordinary GLMs. Our proposal is to extend the Lin *et al.* (2002) methods developed for GLM to ROC regression models by considering cumulative sums of residuals over various coordinates. We derive asymptotic null distributions for the residual processes and show how specific assumptions that pertain to ROC regressions can be examined within this framework. For the first type of aforementioned regression models where Y_{D} and $Y_{\bar{\mathrm{D}}}$ are modelled separately, we develop graphical procedures for checking the adequacy of semi-parametric location-scale models considered by Heagerty and Pepe (1999) and Zheng and Heagerty (2004). For the direct modelling approach, we focus on the semi-parametric ROC-GLM (Cai and Pepe, 2002; Cai, 2004) which extends the parametric ROC-GLM by leaving the parametrically specified baseline function completely unspecified. To examine the aptness of AUC/pAUC regression models, we refer to Cai and Dodd (2004) for detailed model checking procedures.

4

The paper is organized as follows. In Section 2 we outline model checking procedures when FPR and TPR are modelled separately with semi-parametric location-scale models. The procedures are illustrated with a cystic fibrosis dataset. Model checking techniques for semi-parametric ROC-GLMs are presented in Section 3. We apply proposed procedures to examine whether a semi-parametric ROC-GLM is appropriate for the cystic fibrosis data. Simulation results are presented in section 4 to assess the powers of the tests in detecting the model mis-specification while maintaining reasonable sizes. Some closing remarks are given in Section 5.

2. Semi-parametric TPR and FPR Regression Based ROC Model

2.1 Data Structure

We first describe the general data structure considered throughout this paper. Suppose that the data for analysis are organized as N_{D} data records for n_{D} subjects with disease, i.e., $\mathfrak{R}_{\mathrm{D}} = \{(Y_{\mathrm{D}ik}, \mathbf{Z}_{ik}, \mathbf{Z}_{\mathrm{D}ik}),\ k = 1, \ldots, K_i,\ i = 1, \ldots, n_{\mathrm{D}}\}$, and $N_{\bar{\mathrm{D}}}$ data records for $n_{\bar{\mathrm{D}}}$ subjects without disease, i.e., $\mathfrak{R}_{\bar{\mathrm{D}}} = \{(Y_{\bar{\mathrm{D}}jl}, \mathbf{Z}_{jl}),\ l = 1, \ldots, K_j,\ j = n_{\mathrm{D}} + 1, \ldots, n_{\mathrm{D}} + n_{\bar{\mathrm{D}}}\}$. Since each subject may have $K \geq 1$ records, $N_{\mathrm{D}} = \sum_{i=1}^{n_{\mathrm{D}}} K_i$ and $N_{\bar{\mathrm{D}}} = \sum_{j=n_{\mathrm{D}}+1}^{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}} K_j$. The $p_{\bar{\mathrm{D}}} \times 1$ vector $\mathbf{Z}$ denotes covariates relevant to both diseased and non-diseased subjects. Examples include subject characteristic such as age, gender or the type of biomarker represented by Y. The $p \times 1$ vector $\mathbf{Z}_{\mathrm{D}}$ denotes covariates that are solely relevant to diseased subjects. Examples would be the severity of a disease or the time lag between when a test is measured and the onset of the disease. We use $\mathrm{ROC}_{\mathbf{X}}(\cdot)$ to denote the ROC curve that compares the distribution of Y_{D} with covariates $\mathbf{X} = (\mathbf{Z}, \mathbf{Z}_{\mathrm{D}})$ to the distribution of $Y_{\bar{\mathrm{D}}}$ with covariates $\mathbf{Z}$.

5

2.2 Models and Estimators

Zheng and Heagerty (2004) detailed an approach for estimating covariate specific ROC curves by first modelling Y_{D} and $Y_{\bar{\mathrm{D}}}$ separately through semi-parametric location-scale models:

$$Y_{\mathrm{D}ik} = \boldsymbol{\beta}_{\mathrm{D}}^{\mathsf{T}}\vec{\mathbf{X}}_{ik} + e^{\boldsymbol{\alpha}_{\mathrm{D}}^{\mathsf{T}}\vec{\mathbf{X}}_{ik}}\varepsilon_{\mathrm{D}ik}, \tag{1}$$

$$Y_{\bar{\mathrm{D}}jl} = \boldsymbol{\beta}_{\bar{\mathrm{D}}}^{\mathsf{T}}\vec{\mathbf{Z}}_{jl} + e^{\boldsymbol{\alpha}_{\bar{\mathrm{D}}}^{\mathsf{T}}\vec{\mathbf{Z}}_{jl}}\varepsilon_{\bar{\mathrm{D}}jl}, \tag{2}$$

where for any vector $\boldsymbol{a}$, $\vec{\boldsymbol{a}} = (1, \boldsymbol{a})$, the standardized errors $\varepsilon_{\mathrm{D}ik}$ and $\varepsilon_{\bar{\mathrm{D}}jl}$ have conditional mean 0 and variance 1 with unknown distributions. We consider the situation when the errors are independent of covariates. The induced ROC curve for covariate $\mathbf{X} = (\mathbf{Z}, \mathbf{Z}_{\mathrm{D}})$ is

$$\mathrm{ROC}_{\mathbf{X}}(u) = S_{\mathrm{D}}\left\{ e^{\boldsymbol{\alpha}_{\bar{\mathrm{D}}}^{\mathsf{T}}\vec{\mathbf{Z}} - \boldsymbol{\alpha}_{\mathrm{D}}^{\mathsf{T}}\vec{\mathbf{X}}}S_{\bar{\mathrm{D}}}^{-1}(u) + e^{-\boldsymbol{\alpha}_{\mathrm{D}}^{\mathsf{T}}\vec{\mathbf{X}}}(\boldsymbol{\beta}_{\bar{\mathrm{D}}}^{\mathsf{T}}\vec{\mathbf{Z}} - \boldsymbol{\beta}_{\mathrm{D}}^{\mathsf{T}}\vec{\mathbf{X}}) \right\}, \tag{3}$$

where $S_{\mathrm{D}}(y) = \mathrm{P}(\varepsilon_{\mathrm{D}ik} \geq y)$ and $S_{\bar{\mathrm{D}}}(y) = \mathrm{P}(\varepsilon_{\bar{\mathrm{D}}jl} \geq y)$. Heagerty and Pepe (1999) proposed to estimate the regression parameters in (1) and (2) using a quasi-likelihood method and to estimate $S_{\mathrm{D}}(\cdot)$ and $S_{\bar{\mathrm{D}}}(\cdot)$ using the empirical distribution functions of the fitted residuals. To make inference about the induced ROC curve at a given covariate level, Zheng and Heagerty (2004) established the consistency and asymptotic distribution theory for the plug-in estimator of the ROC curve. However, if (1) or (2) is mis-specified, inferences about the underlying covariate effects or covariate specific ROC curves based on these procedures could be invalid. This concern motivates us to consider supplementing the method with model checking procedures.

2.3 Model-Checking Techniques

Since the ROC model in (3) is induced from the TPR and FPR regression models in (1) and (2), we consider assessing the goodness of fit of (3) by checking the aptness of (1) and (2) separately. We will illustrate our method for (2) only since (1) takes the same form.

Model (2) essentially specifies that the distribution of the standardized residual is independent of covariates, i.e. $\mathrm{P}(\varepsilon_{\bar{\mathrm{D}}jl} \geq y \mid \mathbf{Z}_{jl}) = S_{\bar{\mathrm{D}}}(y)$. In other words, the conditional mean of the

6

derived residual, $I(\varepsilon_{\bar{\mathtt{D}}jl} \geq y) - S_{\bar{\mathtt{D}}}(y)$, is 0. We propose to validate this assumption by considering cumulative sums of the derived residuals. The motivation for employing a cumulative residual analysis is that estimation of the conditional distribution of raw residuals requires smoothing and the variability of such an estimator is difficult to assess in general. In contrast, the distribution of the cumulative residuals can be easily approximated by a zero-mean Gaussian process under the null hypothesis of correct model specification. Thus by considering cumulative residuals over different coordinates, corresponding specific assumptions with regard to our semi-parametric ROC models can be tested in an objective fashion. Specifically, to assess the overall goodness-of-fit of model (2), we consider

$$\widehat{M}_{\bar{\mathtt{D}}}(y, \mathbf{z}) = N_{\bar{\mathtt{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathtt{D}}}} \sum_{l=1}^{K_j} I(\mathbf{Z}_{jl} \leq \mathbf{z}) \left\{ I(\widehat{\varepsilon}_{\bar{\mathtt{D}}jl} \geq y) - \widehat{S}_{\bar{\mathtt{D}}}(y) \right\},$$

where $\widehat{\varepsilon}_{\bar{\mathtt{D}}jl} = \{Y_{\bar{\mathtt{D}}jl} - \widehat{\boldsymbol{\beta}}_{\bar{\mathtt{D}}}^{\mathsf{T}} \vec{\mathbf{Z}}_{jl}\} / \exp(\widehat{\boldsymbol{\alpha}}_{\bar{\mathtt{D}}}^{\mathsf{T}} \vec{\mathbf{Z}}_{jl})$ and $\widehat{S}_{\bar{\mathtt{D}}}(y) = N_{\bar{\mathtt{D}}}^{-1} \sum_{j=1}^{n_{\bar{\mathtt{D}}}} \sum_{l=1}^{K_j} I(\widehat{\varepsilon}_{\bar{\mathtt{D}}jl} \geq y)$. Since $E[I(\mathbf{Z}_{jl} \leq \mathbf{z})\{I(\varepsilon_{\bar{\mathtt{D}}jl} \geq y) - S_{\bar{\mathtt{D}}}(y)\}] = 0$ under model (2), we expect $\widehat{M}_{\bar{\mathtt{D}}}(\cdot, \cdot)$ to fluctuate around 0, provided that $\widehat{\boldsymbol{\theta}}_{\bar{\mathtt{D}}} = (\widehat{\boldsymbol{\alpha}}_{\bar{\mathtt{D}}}^{\mathsf{T}}, \widehat{\boldsymbol{\beta}}_{\bar{\mathtt{D}}})^{\mathsf{T}}$ is a consistent estimator for $\boldsymbol{\theta}_{\bar{\mathtt{D}}} = (\boldsymbol{\alpha}_{\bar{\mathtt{D}}}, \boldsymbol{\beta}_{\bar{\mathtt{D}}})$. Thus a large value of $\mathcal{S}_{\bar{\mathtt{D}}} = \sup_{y,\mathbf{z}} \left| \widehat{M}_{\bar{\mathtt{D}}}(y, \mathbf{z}) \right|$ leads to the conclusion of model mis-specification. It is straightforward to show that the test based on $\mathcal{S}_{\bar{\mathtt{D}}}$ is consistent against the general alternative that the location-scale model does not hold. That is, $\mathcal{S}_{\bar{\mathtt{D}}} \to \infty$ as $n \to \infty$ if (2) fails. To assess how unusual the observed statistic $\mathcal{S}_{\bar{\mathtt{D}}}$ is, we need to approximate the null distribution of $\widehat{M}_{\bar{\mathtt{D}}}(y, \mathbf{z})$. Without loss of generality, we assume that $N_{\bar{\mathtt{D}}}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}}_{\bar{\mathtt{D}}} - \boldsymbol{\theta}_{\bar{\mathtt{D}}})$ is asymptotically equivalent to a sum of independent terms, $N_{\bar{\mathtt{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathtt{D}}}} \boldsymbol{\theta}_{\bar{\mathtt{D}}j}$, and converges in distribution to a zero mean multivariate normal. The estimator proposed by Heagerty and Pepe (1999) satisfies this condition. We show in the appendix that the process $\widehat{M}_{\bar{\mathtt{D}}}(\cdot, \cdot)$ is asymptotically equivalent to $\widetilde{M}_{\bar{\mathtt{D}}}(\cdot, \cdot) = N_{\bar{\mathtt{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathtt{D}}}} \mathcal{M}_{\bar{\mathtt{D}}j}(\cdot, \cdot)$ which converges weakly to a zero mean Gaussian process, where

$$\mathcal{M}_{\bar{\mathtt{D}}j}(y, \mathbf{z}) = \sum_{l=1}^{K_j} \{I(\mathbf{Z}_{jl} \leq \mathbf{z}) - H(\mathbf{z})\} \{I(\varepsilon_{\bar{\mathtt{D}}jl} \geq y) - S_{\bar{\mathtt{D}}}(y)\} - \mathcal{B}_1(y, \mathbf{z})^{\mathsf{T}} \mathcal{A}_{\bar{\mathtt{D}}j},$$

7

$H(\mathbf{z}) = P(\mathbf{Z}_{jl} \le \mathbf{z})$, $\mathcal{B}_1(y, \mathbf{z})$ is the limit of $\widehat{\mathcal{B}}_1(y, \mathbf{z})$ and

$$\widehat{\mathcal{B}}_1(y, \mathbf{z}) = N_{\bar{\mathrm{D}}}^{-1} \sum_{j=1}^{n_{\bar{\mathrm{D}}}} \sum_{l=1}^{K_j} f_{\bar{\mathrm{D}}}(y) \left\{ I(\mathbf{Z}_{jl} \le \mathbf{z}) - H(\mathbf{z}) \right\} \begin{pmatrix} y\vec{\mathbf{Z}}_{jl} \\ \vec{\mathbf{Z}}_{jl} e^{-\boldsymbol{\alpha}_{\bar{\mathrm{D}}}^{\mathsf{T}} \vec{\mathbf{z}}_{jl}} \end{pmatrix}.$$

The asymptotic equivalence between $\widehat{M}_{\bar{\mathrm{D}}}(\cdot, \cdot)$ and $\widetilde{M}_{\bar{\mathrm{D}}}(\cdot, \cdot)$ allows us to approximate the limiting distribution of $\widehat{M}_{\bar{\mathrm{D}}}$ using a re-sampling technique (Parzen, Wei and Ying, 1994) in practice. In essence, one can simulate random samples $\mathcal{L} = \{\mathcal{L}_j, j = n_{\mathrm{D}} + 1, n_{\mathrm{D}} + n_{\bar{\mathrm{D}}}\}$ from the standard normal distribution, and for each set of $\mathcal{L}$, compute

$$\widehat{M}_{\bar{\mathrm{D}}}^{\mathcal{L}}(y, \mathbf{z}) = N_{\bar{\mathrm{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathrm{D}}}} \widehat{\mathcal{M}}_{\bar{\mathrm{D}}j}(y, \mathbf{z}) \mathcal{L}_j,$$

where $\widehat{\mathcal{M}}_{\bar{\mathrm{D}}j}(y, z)$ is obtained by replacing all the theoretical quantities in $\mathcal{M}_{\bar{\mathrm{D}}j}(y, \mathbf{z})$ by their empirical counterparts. Using similar arguments as given in appendix, it is straightforward to show that the null distribution of $\widehat{M}_{\bar{\mathrm{D}}}(\cdot, \cdot)$ is the same in the limit as the distribution of $\widehat{M}_{\bar{\mathrm{D}}}^{\mathcal{L}}(\cdot, \cdot)$ conditioning on the observed data. Therefore, realizations of $\widehat{M}_{\bar{\mathrm{D}}}^{\mathcal{L}}(\cdot, \cdot)$ can be used to approximate the asymptotic null distribution of $\widehat{M}_{\bar{\mathrm{D}}}(\cdot, \cdot)$.

To graphically determine whether the observed process deviates from 0, one may plot the observed process $\widehat{M}_{\bar{\mathrm{D}}}(\cdot, \cdot)$ along with a few realizations from $\widehat{M}_{\bar{\mathrm{D}}}^{\mathcal{L}}(\cdot, \cdot)$. A significant difference between the observed pattern and its simulated counterparts from the null distribution would lead to the conclusion of model mis-specification. To enhance the objectivity of this graphical procedure, one may supplement these cumulative residual plots with numerical values. In particular one can report the p-value for testing $\mathrm{H}_{\bar{\mathrm{D}}0}$ against the general alternative. We approximate the p-value by $P(\widehat{\mathcal{S}}_{\bar{\mathrm{D}}}^{\mathcal{L}} \ge s_{\bar{\mathrm{D}}} \mid \mathfrak{R}_{\bar{\mathrm{D}}})$, where $\widehat{\mathcal{S}}_{\bar{\mathrm{D}}}^{\mathcal{L}} = \sup_{y, \mathbf{z}} \left| \widehat{M}_{\bar{\mathrm{D}}}^{\mathcal{L}}(y, \mathbf{z}) \right|$ and $s_{\bar{\mathrm{D}}}$ is the observed value of $\mathcal{S}_{\bar{\mathrm{D}}}$.

However, the lack of fit test based on $\widehat{M}_{\bar{\mathrm{D}}}(y, \mathbf{z})$ is omnibus in nature and hence may not be sensitive to specific model departures. Model (2) has four aspects: (i) the linearity in the mean; (ii) the covariance structure; (iii) the linear functional forms of individual covariates; and (iv) the distribution of the properly standardized residuals being independent of individual covariate. If

8

we are interested in these specific model assumptions, corresponding graphical procedures can be used. We consider

$$\widehat{M}_{\bar{\mathsf{D}}\mu}(z) = N_{\bar{\mathsf{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathsf{D}}}} \sum_{l=1}^{K_j} I(\widehat{\boldsymbol{\beta}}_{\bar{\mathsf{D}}}^{\mathsf{T}} \vec{\mathbf{Z}}_{jl} \le z) \left(Y_{\bar{\mathsf{D}}jl} - \widehat{\boldsymbol{\beta}}_{\bar{\mathsf{D}}}^{\mathsf{T}} \vec{\mathbf{Z}}_{jl} \right)$$

for examining the linearity of the mean;

$$\widehat{M}_{\bar{\mathsf{D}}\sigma}(z) = N_{\bar{\mathsf{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathsf{D}}}} \sum_{l=1}^{K_j} I(\widehat{\boldsymbol{\alpha}}_{\bar{\mathsf{D}}}^{\mathsf{T}} \vec{\mathbf{Z}}_{jl} \le z) \left(\widehat{\varepsilon}_{\bar{\mathsf{D}}jl}^2 - 1 \right)$$

for checking the covariance structure of the residual;

$$\widehat{M}_{\bar{\mathsf{D}}z_q}(z) = N_{\bar{\mathsf{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathsf{D}}}} \sum_{l=1}^{K_j} I(Z_{jlq} \le z) \widehat{\varepsilon}_{\bar{\mathsf{D}}jl}$$

for assessing the functional form of the qth covariate; and

$$\widehat{M}_{\bar{\mathsf{D}}q}(y, z) = N_{\bar{\mathsf{D}}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{\mathsf{D}}}} \sum_{l=1}^{K_j} I(Z_{jlq} \le z) \left\{ I(\widehat{\varepsilon}_{\bar{\mathsf{D}}jl} \ge y) - \widehat{S}_{\bar{\mathsf{D}}}(y) \right\}$$

for examining whether the standardized residuals are independent of the qth covariate. The weak convergence of each of these processes under $\mathrm{H}_{\bar{\mathsf{D}}0}$ can be established with similar arguments used for $\widehat{M}_{\bar{\mathsf{D}}}(\cdot, \cdot)$. Supremum-type statistic can also be constructed to test $\mathrm{H}_{\bar{\mathsf{D}}0}$ against specific alternatives and p-values for these tests can be ascertained accordingly based on their approximated distributions. We note that when there is evidence that the error distribution depends on the covariates, the covariate specific FPR function needs to be estimated using local smoothing methods. See Zheng and Heagerty (2004) for details of such procedures.

2.4   Cystic Fibrosis Data Example

We now illustrate the proposed model checking techniques with the Cystic Fibrosis data studied by Zheng and Heagerty (2004) using semi-parametric location-scale models. This dataset comes from the U.S. Cystic Fibrosis Foundation National Patient Registry which contains longitudinal measures of participant health status. For patients diagnosed with cystic fibrosis, FEV_1, the forced

9

expiratory volume in one second, is considered a predictive marker for disease progression, with higher values of FEV_1 associated with better health. It is of great clinical interest to quantify the degree to which FEV_1 can discriminate between patients who subsequently suffer a pulmonary exacerbation and those who do not. For illustration, we consider a subset of female patients between ages of 6 and 18 years with data recorded in both 1995 and 1996 and who tested negative for pseudomonas aeruginosa with a throat culture. Subjects were considered as diseased if they experienced a pulmonary exacerbation in 1996 and disease-free otherwise. The discriminatory marker is $Y = -FEV_1$ with FEV_1 measured in 1995.

We first examine whether the marker distribution for the $n_{\bar{D}} = 1055$ controls follows a location-scale model:

$$Y_{\bar{D}} = \beta_{\bar{D}0} + \beta_{\bar{D}1}\text{age} + \beta_{\bar{D}2}\text{height} + \exp(\alpha_{\bar{D}0} + \alpha_{\bar{D}1}\text{age} + \alpha_{\bar{D}2}\text{height})\varepsilon_{\bar{D}}. \tag{4}$$

The estimated regression coefficients by fitting the above model are presented in Table 1(a). We apply the model checking procedures for FPR functions to examine if the model for FEV_1 is appropriately specified with respect to 1) the linearity in the mean; 2) the variance structure; 3) the covariate functional forms; and 4) the conditional distribution of the standarized residuals. Figure 1 shows the cumulative residual plots for the fitted FPR model. As shown in Figure 1(a), the cumulative residual plot for the mean suggests that it may not be linear in age or height. This is objectively confirmed with a p-value of 0.001 from the supremum test. In contrast, the cumulative residual plot for the variance structure, shown in Figure 1(b), suggests that the variance structure is reasonably modelled, with a p-value of 0.604 based on the supremum test. To examine the functional forms of age and height, we display the graphs of cumulative residuals versus age (Figure 1(c)) and height(Figure 1(d)) respectively. The residual plots and supremum tests indicate possibly incorrect choice of the linear functional form for age. In addition, from Figure 1(e) and Figure 1(f), it appears that the distribution of the standardized residuals may be dependent on

10

age (p-value = 0.004) but not on height (p-value = 0.918). To account for possible non-linearity in the mean and non-linear age effect, we considered a logarithmic transformation for $Y_{\bar{\mathrm{D}}}$, and added age^2 to the model. Table 1(b) shows the estimated regression coefficients under the new model. The corresponding residual plots, displayed in Figure 2, suggest that no significant model misspecification can now be detected.

Similarly, we applied the model checking procedures to the semi-parametric location-scale model for cases ($n_{\mathrm{D}} = 235$) based on $\log(-\mathrm{FEV}_{\mathrm{1D}}) = \beta_{\mathrm{D0}} + \beta_{\mathrm{D1}}\mathrm{height} + \beta_{\mathrm{D2}}\mathrm{age} + \beta_{\mathrm{D3}}\mathrm{age}^2 + \exp(\alpha_{\mathrm{D0}} + \alpha_{\mathrm{D1}}\mathrm{height} + \alpha_{\mathrm{D2}}\mathrm{age} + \alpha_{\mathrm{D3}}\mathrm{age}^2)\varepsilon_{\mathrm{D}}$. This model appears adequate based on residual plots and supremum tests. To illustrate the impact of model mis-specification on the resulting ROC curves, we show in Figure 3 the estimated ROC curves based on our final models and the original models as given in (4) for the sub-population with covariates fixed at the sample median level, age = 11 and (standardized) height = 0.6. At this covariate level, the mis-specified models result in an over-estimation of the accuracy of FEV1. For example, at FPR = 0.20, the estimated sensitivity is 0.56 based on the corrected models but is about 0.69 based on the mis-specified models.

3. Semi-parametric ROC-GLM Regression Model

3.1 Models and Estimators

In contrast to the approach where Y_{D} and $Y_{\bar{\mathrm{D}}}$ are modelled separately to induce an ROC curve, the direct modelling approach, first proposed by Pepe (1997), specifies the covariate effects on the ROC curve directly through a parametric ROC-GLM:

$$\mathrm{ROC}_{\mathbf{X}_{ik}}(u) = \mathrm{P}\left\{Y_{\mathrm{D}ik} \geq S_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}^{-1}(u) \mid \mathbf{X}_{ik}\right\} = g\left\{h_{\boldsymbol{\alpha}_0}(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\right\}, \tag{5}$$

where $S_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$ is the FPR function associated with covariates $\mathbf{Z}$, $g : (-\infty, \infty) \to (0,1)$ is some pre-specified increasing link function and $h_{\boldsymbol{\alpha}}(\cdot)$ is known up to a finite dimensional parameter vector $\boldsymbol{\alpha}$. In this model, the baseline function $h_{\boldsymbol{\alpha}}$ essentially defines the location and the shape of an ROC curve while $\boldsymbol{\beta}$ quantifies the covariate effect on the diagnostic accuracy. Cai and Pepe

11

(2002) and Cai (2004) extended the parametric ROC-GLM to a semi-parametric ROC-GLM by allowing an arbitrary non-parametric baseline function $h_0(\cdot)$:

$$\mathrm{ROC}_{\mathbf{X}_{ik}}(u) = g\left\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\right\}. \tag{6}$$

Simulation studies suggested that the extra flexibility in the semi-parametric model is gained with little loss in statistical efficiency. One may specific model (6) for u in $(0,1)$ or in a sub-interval of $(0,1)$. Partial ROC regression may be desirable when only a specific range of false positive rates is clinically relevant. To accommodate this, we develop procedures for examining the adequacy of the semi-parametric ROC-GLM for $u \in [a,b] \subset (0,1)$.

To estimate the unknown parameters $\boldsymbol{\psi}_0(\cdot) = \begin{bmatrix} h_0(\cdot) \\ \boldsymbol{\beta}_0 \end{bmatrix}$ under model (6), estimating equations based on generalized method of moments (Cai and Pepe, 2002; Cai, 2004) have been proposed. These inference procedures were derived on the basis of the placement value of Y_{D} (Hanley and Haijian-Tilaki, 1997; Pepe and Cai, 2004), defined as $U_{\mathrm{D}} \equiv S_{\bar{\mathrm{D}},\mathbf{Z}}(Y_{\mathrm{D}})$. The validity of these procedures for making inference about $\mathrm{ROC}_{\mathbf{X}}(u)$ relies on the following conditions: a) $S_{\bar{\mathrm{D}},\mathbf{z}}(\cdot)$ is consistently estimated and b) the ROC-GLM in (6) is correctly specified. To examine the first condition, model checking procedures discussed in section 2 can be used if we consider a semi-parametric location-scale model specified in (2). Therefore in this section we focus on examining the second condition, i.e., whether the ROC-GLM specified in (6) is appropriate for the data.

3.2 Model Checking Techniques

We first consider a global test for the goodness of fit of the model and then propose two tests that are more sensitive in detecting the mis-specification in certain components of the ROC-GLM. Let $\widehat{\boldsymbol{\psi}}(u) = \begin{bmatrix} \widehat{h}(u) \\ \widehat{\boldsymbol{\beta}} \end{bmatrix}$ denotes an estimator of $\boldsymbol{\psi}_0(u)$ that is consistent under model (6). We also assume that the FPR function $S_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$ is consistently estimated by $\widehat{S}_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$.

12

A global test for the ROC-GLM To examine the adequacy of ROC-GLM, we note that $\mathrm{ROC}_{\mathbf{X}_{ik}}(\cdot)$ is the marginal distribution function of the placement value $U_{\mathrm{D}ik} = S_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(Y_{\mathrm{D}ik})$. If model (6) holds only for $u \in [a, b]$, it essentially specifies $\mathrm{P}(U_{\mathrm{D}ik} \leq u \mid U_{\mathrm{D}ik} \in [a, b], \mathbf{X}_{ik})$ as

$$F_{\mathbf{X}_{ik}}(u) \equiv \frac{g\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\} - g\{h_0(a) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\}}{g\{h_0(b) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\} - g\{h_0(a) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik}\}} = \frac{g\{\boldsymbol{\psi}_0(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\} - g\{\boldsymbol{\psi}_0(a)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}}{g\{\boldsymbol{\psi}_0(b)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\} - g\{\boldsymbol{\psi}_0(a)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}}, \qquad u \in [a, b].$$

Thus, conditional on $U_{\mathrm{D}ik} \in [a, b]$, the random variable $V_{\mathrm{D}ik} = F_{\mathbf{X}_{ik}}(U_{\mathrm{D}ik})$ follows a Uniform(0, 1) distribution and is independent of $\mathbf{X}_{ik}$. Consequently, $E[I(U_{\mathrm{D}ik} \in [a, b])\{I(V_{\mathrm{D}ik} \leq u) - u\} \mid \mathbf{X}_{ik}] = 0$. This motivates us to examine the adequacy of model (6) by considering the following cumulative residual process based on $I(\widehat{V}_{\mathrm{D}ik} \leq u) - u$:

$$\widehat{M}(u, \mathbf{x}) = N_{\mathrm{D}}^{-1} \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} \sum_{k=1}^{K_i} I(\mathbf{X}_{ik} \leq \mathbf{x}, \widehat{U}_{\mathrm{D}ik} \in [a, b]) \left\{ I(\widehat{V}_{\mathrm{D}ik} \leq u) - u \right\},$$

where $\widehat{V}_{\mathrm{D}ik} = \widehat{F}_{\mathbf{X}_{ik}}(\widehat{U}_{\mathrm{D}ik})$, $\widehat{F}_{\mathbf{X}}(u)$ is the plug-in estimator of $F_{\mathbf{X}}(u)$ replacing $\boldsymbol{\psi}_0$ by $\widehat{\boldsymbol{\psi}}$ and $\widehat{U}_{\mathrm{D}ik} = \widehat{S}_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(Y_{\mathrm{D}ik})$. Under the null hypothesis that model (6) holds, $\widehat{F}_{\mathbf{X}}(u)$ is consistent and thus we expect that $\widehat{M}(u, \mathbf{x})$ fluctuates around 0. Moreover, we can show that under (6) and mild regularity conditions that the process $N_{\mathrm{D}}^{\frac{1}{2}}\widehat{M}(u, \mathbf{x})$ converges weakly to a zero-mean Gaussian process by establishing the weak convergence of the process $n_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{F}_{\mathbf{X}}(u) - F_{\mathbf{X}}(u)\right\}$ and accounting for the variability in estimating the placement values. Thus, a large value of $\mathcal{S} = \sup_{u,\mathbf{x}} \left|\widehat{M}(u, \mathbf{x})\right|$ leads to the conclusion of model mis-specification. It is straightforward to show that a test based on $\mathcal{S}$ is consistent since $\mathcal{S}$ converges to a positive constant when model (6) does not hold.

A test for link function One main concern in employing a semi-parametric ROC-GLM is in the choice of the link function $g(\cdot)$. To examine the appropriateness of a given link function g, we mimic the cumulative residual process proposed by Lin et al (2002) for examining the link function in the GLM and consider residuals aggregated over the predicted values $\widehat{\boldsymbol{\beta}}^{\mathsf{T}}\mathbf{X}_{ik}$. Instead of examining the entire distribution function of $\widehat{V}_{\mathrm{D}ik}$, we use the mean and the variance of $V_{\mathrm{D}ik}$ as summary

13

indices of the distribution and examine whether $E(V_{\mathrm{D}ik} \mid \mathbf{X}_{ik}) = 1/2$ and $E(V_{\mathrm{D}ik}^2 \mid \mathbf{X}_{ik}) = 1/3$. Specifically, we consider the following cumulative residual process for examining the link function

$$\widehat{\mathbf{M}}_L(x) = N_{\mathrm{D}}^{-1} \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} \sum_{k=1}^{K_i} I(\widehat{\boldsymbol{\beta}}^{\mathsf{T}} \mathbf{X}_{ik} \leq x, a \leq \widehat{U}_{\mathrm{D}ik} \leq b) \begin{bmatrix} \widehat{V}_{\mathrm{D}ik} - \frac{1}{2} \\ \widehat{V}_{\mathrm{D}ik}^2 - \frac{1}{3} \end{bmatrix}. \tag{7}$$

In Appendix B, we show that under correct model specification and mild regularity conditions, the residual process $N_{\mathrm{D}}^{\frac{1}{2}} \widehat{\mathbf{M}}_L(x)$ is asymptotically equivalent to $N_{\mathrm{D}}^{-\frac{1}{2}} \{ \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} \mathbf{M}_{\mathrm{L D}i}(x) + \sum_{j=1}^{n_{\bar{\mathrm{D}}}} \mathbf{M}_{\mathrm{L}\bar{\mathrm{D}}j}(x) \}$ and converges weakly to a zero-mean Gaussian process, where $\mathbf{M}_{\mathrm{L D}i}(x)$ and $\mathbf{M}_{\mathrm{L}\bar{\mathrm{D}}j}(x)$ are defined in Appendix B. The limiting null distribution can be approximated through the perturbation method similar to what is described in section 2. Specifically, one may generate independent standard normal random variables $\mathcal{L} = \{\mathcal{L}_1, ..., \mathcal{L}_{n_{\mathrm{D}}+n_{\bar{\mathrm{D}}}}\}$ and use realizations of $N_{\mathrm{D}}^{\frac{1}{2}} \widehat{\mathbf{M}}_L^{\mathcal{L}}(x) = N_{\mathrm{D}}^{-\frac{1}{2}} \{ \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} \widehat{\mathbf{M}}_{\mathrm{L D}i}(x) \mathcal{L}_i + \sum_{j=1}^{n_{\bar{\mathrm{D}}}} \widehat{\mathbf{M}}_{\mathrm{L}\bar{\mathrm{D}}j}(x) \mathcal{L}_j \}$ to approximate the distribution of $N_{\mathrm{D}}^{\frac{1}{2}} \widehat{\mathbf{M}}_L(x)$, where $\widehat{\mathbf{M}}_{\mathrm{L D}i}(x)$ and $\widehat{\mathbf{M}}_{\mathrm{L}\bar{\mathrm{D}}j}(x)$ are the respective empirical counterparts of $\mathbf{M}_{\mathrm{L D}i}(x)$ and $\mathbf{M}_{\mathrm{L}\bar{\mathrm{D}}j}(x)$. To test the goodness of fit for the specified link function, one may compare the observed pattern of $\widehat{\mathbf{M}}_L(\cdot)$ to realizations from $N_{\mathrm{D}}^{\frac{1}{2}} \widehat{\mathbf{M}}_L^{\mathcal{L}}(x)$ and estimate a p-value of significance by comparing the observed supremum-type statistics, $\mathcal{S}_L = \sup_x \|\widehat{\mathbf{M}}_{\mathrm{L}}(x)/\widehat{\boldsymbol{\sigma}}_{\mathrm{L}}(x)\|$ to its approximated null distribution, where $\widehat{\boldsymbol{\sigma}}_{\mathrm{L}}(x)$ is the estimated standard error vector of $\widehat{\mathbf{M}}_{\mathrm{L}}(x)$ and $\|\cdot\|$ is the sup-norm.

A test for interaction The form of ROC-GLM implies that covariate effects on the ROC curve do not vary with the false positive rate u. The assumption may not hold if there is an interaction between the baseline function and some covariate. The adequacy of such an assumption can be evaluated by considering the following alternative

$$\mathrm{H}_{\mathrm{I}1} \ : \ \mathrm{ROC}_{\mathbf{x}}(u) = g \left\{ h_1(u) + \boldsymbol{\beta}_1(u)^{\mathsf{T}} \mathbf{x} \right\}, \tag{8}$$

for some non-constant $\boldsymbol{\beta}_1(\cdot)$. Testing H_0 against $\mathrm{H}_{\mathrm{I}1}$ can be achieved by deriving an estimator for $\boldsymbol{\beta}_1(\cdot)$ under $\mathrm{H}_{\mathrm{I}1}$ and examining whether $\boldsymbol{\beta}_1(\cdot)$ is constant over $[a, b]$. To this end, we propose to

14

estimate $\boldsymbol{\psi}_1(u) = \begin{bmatrix} h_1(u) \\ \boldsymbol{\beta}_1(u) \end{bmatrix}$ under $\mathrm{H}_{\mathrm{I}1}$ for any given u by solving the marginal normal equation for the binary variable $I(\widehat{U}_{\mathrm{D}ik} \le u)$:

$$\widehat{\boldsymbol{\Xi}}(u, \boldsymbol{\psi}) = N_{\mathrm{D}}^{-1} \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} \sum_{k=1}^{K_i} W\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}\vec{\mathbf{X}}_{ik}\left[I(\widehat{U}_{\mathrm{D}ik} \le u) - g\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}\right] = 0, \qquad (9)$$

where $W(x) = \dot{g}(x)/[g(x)\{1-g(x)\}]$. In Appendix B, we show that under the alternative $\mathrm{H}_{\mathrm{I}1}$, $\widehat{\boldsymbol{\psi}}_1(\cdot)$ is a consistent of $\boldsymbol{\psi}_1(\cdot)$ and the process $N_{\mathrm{D}}^{-\frac{1}{2}}\{\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)\}$ converges weakly to a zero-mean Gaussian process. To assess whether the covariate effect is the same across all false positive rates on $[a,b]$, a natural approach is to examine whether

$$\widehat{\mathbf{M}}_{\mathrm{I}}(u) = N_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{\boldsymbol{\beta}}_1(u) - \widehat{\boldsymbol{\beta}}\right\}$$

fluctuates around 0. We show also in Appendix B that under the null, the process $\widehat{\mathbf{M}}_{\mathrm{I}}(u)$ converges weakly to a zero-mean Gaussian process for $u \in [a,b]$. The null distribution of $\widehat{\mathbf{M}}_{\mathrm{I}}(u)$ can easily be approximated by using the resampling technique. To test whether the effect of the qth covariate on the ROC curve is constant over the range of FPRs, we can compare the observed $\widehat{M}_{\mathrm{I}q}(u)$ to its null distribution and construct a test based on the supremum-type statistic $\mathcal{S}_{\mathrm{I}}^{(q)} = \sup_u |\widehat{M}_{\mathrm{I}q}(u)|$, where $\widehat{M}_{\mathrm{I}q}(u)$ is the qth component of $\widehat{\mathbf{M}}_{\mathrm{I}}(u)$. The corresponding p-values can be estimated based on the asymptotic null distributions derived above.

3.3 Cystic Fibrosis Example

To illustrate the aforementioned model-checking procedure, we examine whether the following semi-parametric ROC-GLM is appropriate for the cystic fibrosis dataset:

$$\mathrm{ROC}_{\mathbf{X}}(u) = \Phi\left\{h(u) + \beta_1\mathrm{Height} + \beta_2\mathrm{Age} + \beta_3\mathrm{Age}^2\right\} \qquad (10)$$

over the range $u \le 0.5$. For the FPR function, we consider the semi-parametric location-scale model for $-\log Y_{\bar{\mathrm{D}}}$ with the same set of covariates as in section 2. We first examine the assumption that the probit link is a reasonable choice by considering the cumulative residual process $\widehat{\mathbf{M}}_{\mathrm{L}}(x)$.

In Figure 4 we compare the observed process to realizations from the null distribution of $\widehat{\mathbf{M}}_{\mathrm{L}}(x)$. The observed residual pattern is similar to that of the realizations from the null distribution with p-value of 0.91 for testing the null based on $\mathcal{S}_{\mathrm{L}}$. Now, to examine whether covariate effects on the ROC curve are constant over the FPRs, we fit the following FPR-varying covariate effect model

$$\mathrm{ROC}_{\mathbf{X}}(u) = \Phi\left\{h(u) + \beta_1(u)\mathrm{Height} + \beta_2(u)\mathrm{Age} + \beta_3(u)\mathrm{Age}^2\right\}.$$

The estimated curves for covariate effects shown in Figure 5(a-c) do not appear to vary with FPRs. To test the hypothesis of constant covariate effects, we plot the observed $\widehat{\mathbf{M}}_{\mathrm{I}}$ along with 10 realizations from its null distribution in Figure 5(d-f) separately for each covariate. The p-values for testing constant covariate effects are 0.70 for Height, 0.43 for Age and 0.42 for Age^2, respectively. Thus we conclude that the semi-parametric ROC-GLM given in (10) is a reasonable choice.

4. Simulation Studies

We conducted simulation studies to examine the empirical size and power of the proposed tests. Throughout, we use a single covariate Z in the FPR model generated from Uniform$(0, 1)$. For the location-scale FPR model, we simulated $Y_{\bar{\mathrm{D}}}$ from

$$Y_{\bar{\mathrm{D}}} = 10 + 2Z + e^{0.25 + \alpha_1 Z}\varepsilon_{\bar{\mathrm{D}}}, \quad \varepsilon_{\bar{\mathrm{D}}} \sim N(0, 1),$$

for both $\alpha_1 = 0$ and $\alpha_1 = 0.2$. The empirical sizes of proposed tests range from 4.3% to 5.6% at a sample size of 200 with a significance level of 5%. Empirically, we also find that the proposed tests have reasonable power in detecting the mis-specification of the location-scale model. For example, the validity of the derived ROC curve (3) requires the assumption that the residual $\varepsilon_{\bar{\mathrm{D}}}$ is independent of the covariates. To examine the power of the supremum-type test based on $\widehat{M}_{\bar{\mathrm{D}}q}(\cdot, \cdot)$ in detecting the dependence of $\varepsilon_{\bar{\mathrm{D}}}$ on covariates, we simulated non-diseased data with a single covariate Z based on the above location-scale model with $\alpha_1 = 0.2$ but let $\varepsilon_{\bar{\mathrm{D}}}$ be the

16

product of two random variables: one from a Gamma distribution with both shape and rate parameter set at e^{-Z}, the other from a normal distribution with mean 0 and variance $1 + e^Z$. Under this configuration, the power of the test in detecting the dependence of $\varepsilon_{\bar{\text{D}}}$ on Z based on $\sup_{y,z} |\widehat{M}_{\bar{\text{D}}_q}(y,z)|$ is about 75% at a sample size of 200.

For the ROC-GLM, we simulated non-diseased data from a location scale model:

$$Y_{\bar{\text{D}}} = 10 + Z + \varepsilon_{\bar{\text{D}}}, \quad \varepsilon_{\bar{\text{D}}} \sim N(0,1).$$

To examine the empirical size of the tests, we first generated diseased data from

$$Y_{\text{D}} = 10.5 + Z + 0.5Z_{\text{D}} + \varepsilon_{\text{D}}, \quad \varepsilon_{\text{D}} \sim N(0,1),$$

with the additional covariate $Z_{\text{D}} \sim N(0,1)$. This configuration induces an ROC curve of the form: $\text{ROC}_{Z,Z_{\text{D}}}(u) = \Phi\left\{0.5 + 0.5\Phi^{-1}(u) + 0.5Z_{\text{D}}\right\}$. When $n_{\text{D}} = n_{\bar{\text{D}}} = 200$, $a = 0.05$ and $b = 0.95$, the empirical type I error is 4.2% for $\mathcal{S}_{\text{L}}$, 5.5% for $\mathcal{S}_I^{(1)}$ and 4.6% for $\mathcal{S}_I^{(2)}$. We then examine the power of the test based on $\widehat{\mathbf{M}}_{\text{L}}(\cdot)$ in detecting the mis-specification in the link function. In practice, one may expect that the disease population is heterogeneous, especially for complex multistage diseases such as cancer. We therefore simulated Y_{D} from a normal mixture

$$Y_{\text{D}} = 11 + Z + Z_{\text{D}} + \varepsilon_{\text{D}}, \quad \varepsilon_{\text{D}} \sim \left\{ \begin{array}{ll} N(0,1) & \text{with probability } 0.1 \\ N(1,0.5^2) & \text{with probability } 0.9 \end{array} \right. .$$

The resulting ROC curve follows the form of a semi-parametric ROC-GLM but with a link function different from $\Phi(\cdot)$. We find in this case the power is about 90% based on $\mathcal{S}_{\text{L}}$ when the link function is mis-specified as $\Phi(\cdot)$. To examine the power for $\mathcal{S}_I^{(q)}$, we consider a setting where $Z_{\text{D}} \sim \text{Bernoulli}(0.5)$ and generated Y_{D} such that the the true ROC curve is

$$\text{ROC}_{Z,Z_{\text{D}}}(u) = \Phi\left[0.5 + 0.5\Phi^{-1}(u) + \{0.5 + 0.5\Phi^{-1}(u)\}Z_{\text{D}}\right] \tag{11}$$

which has different shape for $Z_{\text{D}} = 0$ and $Z_{\text{D}} = 1$. In practice Z_{D} may represent disease severity and the accuracy of a marker in detecting disease may follow different models depending on Z_{D}. It

17

is tempting to fit a semi-parametric ROC-GLM of the form $\mathrm{ROC}_{Z,Z_{\mathrm{D}}}(u) = \Phi\left\{h(u) + \beta_1 Z + \beta_2 Z_{\mathrm{D}}\right\}$ so that β_2 directly quantifies the effect of disease severity Z_{D} on the accuracy. However prior to employing such an model, it is crucial to check that the ROC curves for different disease severities are indeed have the same shapes as the model assumed. We find that under the model given in (11), the proposed test has high power in detecting the non-constant covariate effect of Z_{D}. For example, at a sample size of $n_{\bar{\mathrm{D}}} = n_{\mathrm{D}} = 200$ and for $u \leq 0.5$, the empirical power of $\mathcal{S}_I^{(2)}$ is about 91% while the empirical power of $\mathcal{S}_I^{(1)}$ remains 4.8%.

5. Remarks

Compared with generalized linear model where attention is often directed towards the mean function of a single population, ROC regression also involves simultaneously evaluating the test result distributions in both diseased and non-diseased populations and the *relationship* between them. The increased complexity in both estimation and inference procedures of ROC regression models makes existing model checking methods less tenable. Therefore there is a need for rigorous and practical procedures to guide the ROC regression modelling in practice. In this paper we address this need by developing model checking procedures for some commonly used ROC regression models. The proposed procedures have the advantage of being objective and are easy to implement in practice. Asymptotic null distributions for the proposed residual processes are developed and a resampling based method is provided to approximate the limiting null distributions. Our numerical studies suggest that the proposed tests have good power in detecting the corresponding model mis-specifications while at the same time maintaining adequate size of the tests. In addition, our proposed methods allow for correlated observations which arise often in medical studies.

The form of the ROC-GLM, $\mathrm{ROC}_{\mathbf{X}}(u) = g\{h_0(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}\}$, suggests that the covariate effect is the same over all false positive rates u. This assumption may not hold in practice, especially if a marker has very different distributions among different sub-populations. Thus the effect of

18

a certain covariate on the marker accuracy may vary with the false positive rate and under such setting it may be preferable to model the ROC curve using the FPR-varying covariate effect model (8). As part of our model checking proposals, we provided procedures for making inference about the FPR-specific covariate effect and the corresponding ROC curve under this model.

Appendix
A.   Null Distributions of $\widehat{M}_{\bar{D}}(y, \mathbf{z})$

For technical reasons, we assume that potentially each subject has $\max(K_1, ..., K_{n_{\bar{D}}})$ records if non-diseased and the $n_{\bar{D}}$ sets of clustered observations $\{(Y_{\bar{D}}, \mathbf{Z})\}$ are independent and identically distributed. Although not every non-diseased subject has $K_{\bar{D}}$ records, the presence or absence of individual records in a cluster does not depend on the observations. Corresponding assumptions are made for observations $\{(Y_D, \mathbf{Z}, \mathbf{Z}_D)\}$ from the diseased subjects. We also assume that the covariates are bounded and $S_{\bar{D}}(\cdot)$ is twice continuously differentiable.

The uniform consistency of $N_{\bar{D}}^{-\frac{1}{2}} \widehat{M}_{\bar{D}}(y, \mathbf{z}) \to 0$ follows from the consistency of $\widehat{\boldsymbol{\theta}}_{\bar{D}}$ and the uniform convergence of $N_{\bar{D}}^{-1} \sum_{j,l} I(\varepsilon_{\bar{D}jl} \geq y) \to S_{\bar{D}}(y)$ in probability, where $\sum_{j,l}$ denotes the summation $\sum_{j=1}^{n_{\bar{D}}} \sum_{l=1}^{K_j}$. To show the weak convergence of $\widehat{M}_{\bar{D}}(y, \mathbf{z})$, we note that by the standard empirical process theory (Pollard, 1990), $\widehat{\eta}(y, \mathbf{z}, \boldsymbol{a}, \mathbf{b}) = N_{\bar{D}}^{-\frac{1}{2}} \sum_{j,l} w_{jl}(\mathbf{z})[I\{\varepsilon_{\bar{D}jl} \geq c_{\mathbf{Z}_{jl}}(y, \boldsymbol{a}, \mathbf{b})\} - S_{\bar{D}}\{c_{\mathbf{Z}_{jl}}(y, \boldsymbol{a}, \mathbf{b})\}]$ converges weakly to a Gaussian and is stochastic equi-continuous, where $w_{jl}(\mathbf{z}) = I(\mathbf{Z}_{jl} \leq \mathbf{z}) - H(\mathbf{z})$ and $c_{\mathbf{Z}}(y, \boldsymbol{a}, \mathbf{b}) = y e^{\boldsymbol{a}^{\mathsf{T}} \mathbf{Z}} + \mathbf{b}^{\mathsf{T}} \mathbf{Z} e^{-\boldsymbol{\alpha}_{\bar{D}}^{\mathsf{T}} \mathbf{Z}}$. It follows that

$$\widehat{M}_{\bar{D}}(y, \mathbf{z}) \simeq \widehat{\eta}(y, \mathbf{z}, \widehat{\boldsymbol{\alpha}}_{\bar{D}} - \boldsymbol{\alpha}_{\bar{D}}, \widehat{\boldsymbol{\beta}}_{\bar{D}} - \boldsymbol{\beta}_{\bar{D}}) + N_{\bar{D}}^{-\frac{1}{2}} \sum_{j,l} w_{jl}(\mathbf{z}) \left[S_{\bar{D}} \left\{ c_{\mathbf{Z}_{jl}}(y, \widehat{\boldsymbol{\alpha}}_{\bar{D}} - \boldsymbol{\alpha}_{\bar{D}}, \widehat{\boldsymbol{\beta}}_{\bar{D}} - \boldsymbol{\beta}_{\bar{D}}) \right\} - S_{\bar{D}}(y) \right]$$

$$\simeq \widehat{\eta}(y, \mathbf{z}, \mathbf{0}, \mathbf{0}) - N_{\bar{D}}^{-\frac{1}{2}} \sum_{j,l} f_{\bar{D}}(y) w_{jl}(\mathbf{z}) \left\{ c_{\mathbf{Z}_{jl}}(y, \widehat{\boldsymbol{\alpha}}_{\bar{D}} - \boldsymbol{\alpha}_{\bar{D}}, \widehat{\boldsymbol{\beta}}_{\bar{D}} - \boldsymbol{\beta}_{\bar{D}}) - y \right\}$$

$$\simeq N_{\bar{D}}^{-\frac{1}{2}} \sum_{j,l} w_{jl}(\mathbf{z}) \left[I\left(\varepsilon_{\bar{D}jl} \geq y \right) - S_{\bar{D}}(y) \right] - N_{\bar{D}}^{\frac{1}{2}} (\widehat{\boldsymbol{\theta}}_{\bar{D}} - \boldsymbol{\theta}_{\bar{D}})^{\mathsf{T}} \mathcal{B}_1(y, \mathbf{z}) = N_{\bar{D}}^{-\frac{1}{2}} \sum_{j=1}^{n_{\bar{D}}} \mathcal{M}_{\bar{D}j}(y, \mathbf{z})$$

Then by the functional central limit theorem (Pollard, 1990), $\widehat{M}_{\bar{D}}(y, \mathbf{z})$ converges weakly to a zero-mean Gaussian process.

19

B. Asymptotic Null Distributions for $\widehat{\mathbf{M}}_{\mathrm{L}}(\cdot)$ and $\widehat{\mathbf{M}}_{\mathrm{I}}(\cdot)$

To derive the null distributions of $\widehat{\mathbf{M}}_{\mathrm{L}}(\cdot)$ and $\widehat{\mathbf{M}}_{\mathrm{I}}(\cdot)$, we assume a location-scale model (2) for $S_{\bar{\mathrm{D}},\mathbf{Z}}(y)$ and the same regularity conditions given in Appendix A. In addition, we assume that the link function g and the baseline function h are both twice continuously differentiable. We first obtain a large sample approximation for $\widehat{\mathcal{I}}_{\bar{\mathrm{D}},\mathbf{Z}}(u) = S_{\bar{\mathrm{D}},\mathbf{Z}}\left\{\widehat{S}_{\bar{\mathrm{D}},\mathbf{Z}}^{-1}(u)\right\}$ which accounts for the variability due to estimating the placement values. Under (2), it follows from the consistency of $\widehat{\boldsymbol{\beta}}$ and $\widehat{S}_{\bar{\mathrm{D}}}(\cdot)$ that $\widehat{\mathcal{I}}_{\bar{\mathrm{D}},\mathbf{Z}}(u) = S_{\bar{\mathrm{D}}}\left\{\widehat{S}_{\bar{\mathrm{D}}}^{-1}(u)e^{(\widehat{\boldsymbol{\alpha}}_{\bar{\mathrm{D}}} - \boldsymbol{\alpha}_{\bar{\mathrm{D}}})^{\mathsf{T}}\vec{\mathbf{Z}}} + e^{\boldsymbol{\alpha}_{\bar{\mathrm{D}}}^{\mathsf{T}}\vec{\mathbf{Z}}}(\widehat{\boldsymbol{\beta}}_{\bar{\mathrm{D}}} - \boldsymbol{\beta}_{\bar{\mathrm{D}}})^{\mathsf{T}}\vec{\mathbf{Z}}\right\} \to u$ in probability, uniformly in $u \in [a,b]$. Furthermore, using the equi-continuity of the process $\widehat{\eta}(y,\mathbf{z},\boldsymbol{a},\mathbf{b})$ and a functional central limit theorem, we can show that $N_{\bar{\mathrm{D}}}^{\frac{1}{2}}\left\{\widehat{\mathcal{I}}_{\bar{\mathrm{D}},\mathbf{Z}}(u) - u\right\}$ is asymptotically equivalent to $N_{\bar{\mathrm{D}}}^{-\frac{1}{2}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}} I_{\bar{\mathrm{D}}_j}(u,\mathbf{Z})$ and converges weakly to a zero-mean Gaussian process, where $I_{\bar{\mathrm{D}}_j}(u,\mathbf{Z}) = -\sum_{l=1}^{K_j}\left\{I\left(\epsilon_{\bar{\mathrm{D}}jl} \geq S_{\bar{\mathrm{D}}}^{-1}(u)\right) - u\right\} - \{\mathcal{C}_{\bar{\mathrm{D}},\mathbf{Z}}(u) - \mathcal{C}_{\bar{\mathrm{D}}}(u)\}^{\mathsf{T}}\boldsymbol{\theta}_{\bar{\mathrm{D}}_j}$, $\mathcal{C}_{\bar{\mathrm{D}}}(u)$ is the limit of $N_{\bar{\mathrm{D}}}^{-1}\sum_{j,l}\mathcal{C}_{\bar{\mathrm{D}},\mathbf{Z}_{jl}}(u)$ and $\mathcal{C}_{\bar{\mathrm{D}},\mathbf{Z}}(u) = f_{\bar{\mathrm{D}}}(S_{\bar{\mathrm{D}}}^{-1}(u))[S_{\bar{\mathrm{D}}}^{-1}(u)\vec{\mathbf{Z}}^{\mathsf{T}}, e^{-\boldsymbol{\alpha}_{\bar{\mathrm{D}}}^{\mathsf{T}}\vec{\mathbf{Z}}}\vec{\mathbf{Z}}^{\mathsf{T}}]^{\mathsf{T}}$.

Next, we show that for a uniformly bounded function $w(u,\mathbf{X})$, the process

$$
\begin{aligned}
\mathcal{U}^{(w)}(u,\mathbf{x}) &= N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k} w(u,\mathbf{X}_{ik})I(\mathbf{X}_{ik} \leq \mathbf{x})\{I(\widehat{U}_{\mathrm{D}ik} \leq u) - I(U_{\mathrm{D}ik} \leq u)\} \\
&= N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k} w(u,\mathbf{X}_{ik})I(\mathbf{X}_{ik} \leq \mathbf{x})\left[I\{U_{\mathrm{D}ik} \leq \widehat{\mathcal{I}}_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(u)\} - I(U_{\mathrm{D}ik} \leq u)\right]
\end{aligned}
$$

converges weakly to a zero-mean Gaussian process. It follows from the equi-continuity of the process $N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k} w(u,\mathbf{X}_{ik})I(\mathbf{X}_{ik} \leq \mathbf{x})[I\{U_{\mathrm{D}ik} \leq S_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(y)\} - \mathrm{ROC}_{\mathbf{X}_{ik}}\{S_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(y)\}]$ that

$$
\mathcal{U}^{(w)}(u,\mathbf{x}) \simeq N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k} w(u,\mathbf{X}_{ik})I(\mathbf{X}_{ik} \leq \mathbf{x})\dot{\mathrm{ROC}}_{\mathbf{X}_{ik}}(u)\left\{\widehat{\mathcal{I}}_{\bar{\mathrm{D}},\mathbf{Z}_{ik}}(u) - u\right\} \simeq N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\mathcal{U}_{\bar{\mathrm{D}}j}^{(w)}(u,\mathbf{x}),
$$

where $\dot{\mathrm{ROC}}_{\mathbf{x}}(u) = \partial\mathrm{ROC}_{\mathbf{x}}(u)/\partial u$, $\mathcal{U}_{\bar{\mathrm{D}}j}^{(w)}(u,\mathbf{x}) = p_{10}\int_{\mathbf{X} \leq \mathbf{x}} I_{\bar{\mathrm{D}}_j}(u,\mathbf{Z})w(u,\mathbf{X})\dot{\mathrm{ROC}}_{\mathbf{X}}(u)f(\mathbf{X})d\mathbf{X}$, p_{10} is the limit of $N_{\mathrm{D}}/N_{\bar{\mathrm{D}}}$ and $f_X(\cdot)$ is the density function of $\mathbf{X}_{ik}$. With a functional central limit theorem, we can then establish the weak convergence of $\mathcal{U}^{(w)}(u,\mathbf{x})$.

Without loss of generality, we also assume that $N_{\mathrm{D}}^{\frac{1}{2}}\{\widehat{\boldsymbol{\psi}}(u) - \boldsymbol{\psi}_0(u)\}$ is asymptotically equivalent to a sum of i.i.d, $N_{\mathrm{D}}^{-\frac{1}{2}}\{\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\boldsymbol{\psi}_{\bar{\mathrm{D}}_j}(u) + \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}}\boldsymbol{\psi}_{\mathrm{D}_i}(u)\}$, and converges weakly to a zero-

20

mean Gaussian process. To derive the null distribution for the residual processes, we note that $N_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{F}_{\mathbf{X}}(u) - F_{\mathbf{X}}(u)\right\}$ is asymptotically equivalent to

$$N_{\mathrm{D}}^{-\frac{1}{2}}\frac{\dot{g}\{\boldsymbol{\psi}_0(u)^{\mathsf{T}}\vec{\mathbf{X}}\}}{p_{\mathbf{X}}}\left\{\widehat{\boldsymbol{\psi}}(u) - \boldsymbol{\psi}_0(u)\right\} - N_{\mathrm{D}}^{-\frac{1}{2}}\frac{\dot{g}\{\boldsymbol{\psi}_0(a)^{\mathsf{T}}\vec{\mathbf{X}}\}}{p_{\mathbf{X}}}\{1 - F_{\mathbf{X}}(u)\}\left\{\widehat{\boldsymbol{\psi}}(a) - \boldsymbol{\psi}_0(a)\right\} -$$

$$N_{\mathrm{D}}^{-\frac{1}{2}}\frac{\dot{g}\{\boldsymbol{\psi}_0(b)^{\mathsf{T}}\vec{\mathbf{X}}\}}{p_{\mathbf{X}}}F_{\mathbf{X}}(u)\left\{\widehat{\boldsymbol{\psi}}(b) - \boldsymbol{\psi}_0(b)\right\}$$

and is asymptotically equivalent to a sum of i.i.d: $N_{\mathrm{D}}^{-\frac{1}{2}}\left\{\sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} F_{\mathrm{D}i}(u,\mathbf{X}) + \sum_{j=1}^{n_{\bar{\mathrm{D}}}} F_{\bar{\mathrm{D}}j}(u,\mathbf{X})\right\}$, where $p_{\mathbf{X}} = g\{\boldsymbol{\psi}_0(b)^{\mathsf{T}}\vec{\mathbf{X}}\} - g\{\boldsymbol{\psi}_0(a)^{\mathsf{T}}\vec{\mathbf{X}}\}$. process. This, coupled with the large sample properties of the process $\mathcal{U}^{(w)}(u,\mathbf{x})$, implies that

$$N_{\mathrm{D}}^{\frac{1}{2}}\widehat{\mathbf{M}}_{\mathrm{L}}(x) \simeq N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k} I(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik} \leq x, U_{\mathrm{D}ik} \in [a,b])\begin{bmatrix} V_{\mathrm{D}ik} - \frac{1}{2} \\ V_{\mathrm{D}ik}^2 - \frac{1}{3} \end{bmatrix}$$

$$+ \int_{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X} \leq x}\int_a^b \begin{bmatrix} F_{\mathbf{X}}(u) - \frac{1}{2} \\ F_{\mathbf{X}}(u)^2 - \frac{1}{3} \end{bmatrix} d\left[N_{\mathrm{D}}^{-1}\sum_{i,k} I(\mathbf{X}_{ik} \leq \mathbf{X})\{I(\widehat{U}_{\mathrm{D}ik} \leq u) - I(U_{\mathrm{D}ik} \leq u)\}\right]$$

$$+ \int_{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X} \leq x}\int_a^b \begin{bmatrix} \widehat{F}_{\mathbf{X}}(u) - F_{\mathbf{X}}(u) \\ \widehat{F}_{\mathbf{X}}(u)^2 - F_{\mathbf{X}}(u)^2 \end{bmatrix} d\left[N_{\mathrm{D}}^{-1}\sum_{i,k} I(\mathbf{X}_{ik} \leq \mathbf{X}, U_{\mathrm{D}ik} \leq u)\right]$$

$$\simeq N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k} I(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik} \leq x, U_{\mathrm{D}ik} \in [a,b])\begin{bmatrix} V_{\mathrm{D}ik} - \frac{1}{2} \\ V_{\mathrm{D}ik}^2 - \frac{1}{3} \end{bmatrix} + \int_{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X} \leq x}\int_{u \in [a,b]} \begin{bmatrix} F_{\mathbf{X}}(u) - \frac{1}{2} \\ F_{\mathbf{X}}(u)^2 - \frac{1}{3} \end{bmatrix} d\mathcal{U}^{(w_0)}(u,\mathbf{X}) +$$

$$\int_{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X} \leq x}\int_a^b N_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{F}_{\mathbf{X}}(u) - F_{\mathbf{X}}(u)\right\}\begin{bmatrix} 1 \\ 2F_{\mathbf{X}}(u) \end{bmatrix}\dot{\mathrm{ROC}}_{\mathbf{X}}(u)f_X(\mathbf{X})dud\mathbf{X}$$

and thus $N_{\mathrm{D}}^{\frac{1}{2}}\widehat{\mathbf{M}}_{\mathrm{L}}(x) \simeq N_{\mathrm{D}}^{-\frac{1}{2}}\left\{\sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}} \mathbf{M}_{\mathrm{L}\mathrm{D}i}(x) + \sum_{j=1}^{n_{\bar{\mathrm{D}}}} \mathbf{M}_{\mathrm{L}\bar{\mathrm{D}}j}(x)\right\}$, where $w_0(u,\mathbf{X}) = 1$,

$$\mathbf{M}_{\mathrm{L}\mathrm{D}i}(x) = \sum_{k=1}^{K_i} I(\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X}_{ik} \leq x, U_{\mathrm{D}ik} \in [a,b])\begin{bmatrix} V_{\mathrm{D}ik} - \frac{1}{2} \\ V_{\mathrm{D}ik}^2 - \frac{1}{3} \end{bmatrix}$$

$$+ \int_{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X} \leq x}\int_a^b F_{\mathrm{D}i}(u,\mathbf{X})\begin{bmatrix} 1 \\ 2F_{\mathbf{X}}(u) \end{bmatrix}\dot{\mathrm{ROC}}_{\mathbf{X}}(u)f_X(\mathbf{X})dud\mathbf{X},$$

$$\mathbf{M}_{\mathrm{L}\bar{\mathrm{D}}j}(x) = \int_{\boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{X} \leq x}\int_a^b \left\{F_{\bar{\mathrm{D}}j}(u,\mathbf{X})\begin{bmatrix} 1 \\ 2F_{\mathbf{X}}(u) \end{bmatrix}\dot{\mathrm{ROC}}_{\mathbf{X}}(u)f_X(\mathbf{X})dud\mathbf{X} + \begin{bmatrix} F_{\mathbf{X}}(u) - \frac{1}{2} \\ F_{\mathbf{X}}(u)^2 - \frac{1}{3} \end{bmatrix}d\mathcal{U}_{\bar{\mathrm{D}}j}^{(w_0)}(u,\mathbf{X})\right\}.$$

It follows from the functional central limit theorem (Pollard, 1990) that $N_{\mathrm{D}}^{\frac{1}{2}}\widehat{\mathbf{M}}_{\mathrm{L}}(x)$ converges weakly to a zero-mean Gaussian process.

To derive the null distribution of $\widehat{\mathbf{M}}_{\mathrm{I}}$, we first show the asymptotic properties for $\widehat{\boldsymbol{\psi}}_1$ under the alternative model given in $\mathrm{H}_{\mathrm{I}1}$. To this end, we assume that $\boldsymbol{\psi}_1(u)$ is continuously differentiable with bounded derivatives for $u \in [a, b]$ and the parameter space for $\boldsymbol{\psi}(\cdot)$, $\Omega_{\boldsymbol{\psi}}$, is compact. First, it follows from a uniform law of large numbers and the uniform consistency of $\widehat{S}_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$ that

$$\sup_{u,\boldsymbol{\psi}} |\widehat{\boldsymbol{\Xi}}(u,\boldsymbol{\psi}) - E\{\boldsymbol{\Xi}(u,\boldsymbol{\psi})\}| \to 0 \qquad \text{almost surely.} \tag{12}$$

This, together with the consistency of MLE under binary regression models, and uniform convergence of $\boldsymbol{\Xi}(u,\boldsymbol{\psi}) \to E\{\boldsymbol{\Xi}(u,\boldsymbol{\psi})\}$, implies that $\sup_u |\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)|$ almost surely under $\mathrm{H}_{\mathrm{I}1}$. In addition, $\dot{\widehat{\boldsymbol{\Xi}}}\{u, \boldsymbol{\psi}_1(u)\}$ converges almost surely to $\mathbb{A}_1(u) = E[\vec{\mathbf{X}}_{ik}\vec{\mathbf{X}}_{ik}^{\mathsf{T}}W\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}\dot{g}\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}]$, uniformly in u. Therefore $\boldsymbol{\psi}_1$ is the unique solution to $E\{\boldsymbol{\Xi}(u,\boldsymbol{\psi})\} = 0$. This, combined with (12), implies that $\sup_u |\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)|$ almost surely under $\mathrm{H}_{\mathrm{I}1}$. To derive an asymptotic linear expansion for $\widehat{\boldsymbol{\psi}}_1(\cdot)$, we note that $0 = \widehat{\boldsymbol{\Xi}}\{u, \widehat{\boldsymbol{\psi}}_1(u)\} = \widehat{\boldsymbol{\Xi}}\{u, \boldsymbol{\psi}_1(u)\} + \mathbb{A}_1(u)\{\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)\} + O_p(\|\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)\|^2)$ and thus

$$N_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)\right\} \simeq \mathbb{A}_1(u)^{-1}N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k}W\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}\vec{\mathbf{X}}_{ik}\{I(U_{\mathrm{D}ik} \le u) - g(\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik})\}$$
$$+ N_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i,k}\mathbb{A}_1(u)^{-1}W\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}\vec{\mathbf{X}}_{ik}\{I(\widehat{U}_{\mathrm{D}ik} \le u) - I(U_{\mathrm{D}ik} \le u)\}$$

Letting $w_1(u, \mathbf{X}) = \mathbb{A}_1(u)^{-1}W\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}\}\vec{\mathbf{X}}$ and using the asymptotic approximation for $\mathcal{U}^{(w_1)}(u, \infty)$, we obtain the following approximation:

$$N_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{\boldsymbol{\psi}}_1(u) - \boldsymbol{\psi}_1(u)\right\} \simeq N_{\mathrm{D}}^{-\frac{1}{2}}\left\{\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\boldsymbol{\psi}_{1\bar{\mathrm{D}}_j}(u) + \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}}\boldsymbol{\psi}_{1\mathrm{D}_i}(u)\right\} \tag{13}$$

where $\boldsymbol{\psi}_{1\bar{\mathrm{D}}_j}(u) = \mathcal{U}_{\bar{\mathrm{D}}j}^{(w_1)}(u, \infty)$ and $\boldsymbol{\psi}_{1\mathrm{D}_i}(u) = \sum_{k=1}^{K_i}W\{\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik}\}\vec{\mathbf{X}}_{ik}\{I(U_{\mathrm{D}ik} \le u) - g(\boldsymbol{\psi}_1(u)^{\mathsf{T}}\vec{\mathbf{X}}_{ik})\}$. It follows that under $\mathrm{H}_{\mathrm{I}0}$,

$$\widehat{\mathbf{M}}_{\mathrm{I}}(u) \simeq N_{\mathrm{D}}^{\frac{1}{2}}\left[\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\{\boldsymbol{\beta}_{\bar{\mathrm{D}}_1j}(u) - \boldsymbol{\beta}_{\bar{\mathrm{D}}_0j}\} + \sum_{i=n_{\bar{\mathrm{D}}}+1}^{n_{\bar{\mathrm{D}}}+n_{\mathrm{D}}}\{\boldsymbol{\beta}_{\mathrm{D}_1i}(u) - \boldsymbol{\beta}_{\mathrm{D}_0i}\}\right]$$

and $\widehat{\mathbf{M}}_{\mathrm{I}}(u)$ converges weakly to a zero-mean Gaussian process, where $\boldsymbol{\beta}_{\bar{\mathrm{D}}_1j}(u)$, $\boldsymbol{\beta}_{\bar{\mathrm{D}}_0j}$, $\boldsymbol{\beta}_{\mathrm{D}_1i}(u)$ and $\boldsymbol{\beta}_{\mathrm{D}_0i}$ are the $[2:(p_{\mathrm{D}}+1)]$th elements of $\boldsymbol{\psi}_{\bar{\mathrm{D}}_1j}(u)$, $\boldsymbol{\psi}_{\bar{\mathrm{D}}_0j}(u)$, $\boldsymbol{\psi}_{\mathrm{D}_1i}(u)$ and $\boldsymbol{\psi}_{\mathrm{D}_0i}(u)$, respectively.

22

REFERENCES

Alonzo, T. A. and Pepe, M. S. (2002). Distribution-free analysis using binary regression techniques. *Biostatistics* **3**, 421–32.

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

Cai, T. and Dodd, L. (2004). Regression analysis for the partial area under the ROC curve, *submitted*.

Cai, T. and Pepe, M. S. (2002). Semiparametric Receiver Operating Characteristic analysis to evaluate biomarkers for disease. *Journal of the American Statistical Association* **97**, 1099–1107.

Cook, R. D. and Weisberg, S. (1994). *An introduction to regression graphics*. John Wiley & Sons.

Cook, R. D. and Weisberg, S. (1997). Graphics for assessing the adequacy of regression models. *Journal of the American Statistical Association* **92**, 490–499.

Dodd, L. and Pepe, M. S. (2003a). Partial AUC estimation and regression. *Biometrics* **59**, 614–23.

Dodd, L. and Pepe, M. S. (2003b). Semi-parametric regression for the area under the Receiver Operating Characteristic curve. *Journal of the American Statistical Association* **98**, 409–17.

Dorfman, D. and Alf, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology* **6**, 487–496.

Fienberg, S. E. and Gong, G. D. (1984). Comments on "Graphical methods for assessing logistic regression models". *Journal of the American Statistical Association* **79**, 72–77.

Hanley, J. and 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.

Heagerty, P. J. and Pepe, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics* **48**,

23

533–551.

Ismirlian, G. (2003). A new efficient semi-parametric family of models for the regression analysis of ROC curves, *submitted.*

Landwehr, J. M., Pregibon, D. and Shoemaker, A. C. (1984). Graphical methods for assessing logistic regression models (C/R: p72-83). *Journal of the American Statistical Association* **79**, 61–71.

Lin, D. Y., Wei, L. J. and Ying, Z. (2002). Model-checking techniques based on cumulative residuals. *Biometrics* **58**, 1–12.

Obchowski, N. (1995). Multireader Receiver Operating Characteristic studies - a comparison of study designs. *Academic Radiology* **2**, 709–16.

Parzen, M. I., Wei, L. J. and Ying, Z. (1994). A resampling method based on pivotal estimating functions. *Biometrika* **81**, 341–350.

Pepe, M. S. (1997). A regression modelling framework for Receiver Operating Characteristic curves in medical diagnostic testing. *Biometrika* **84**, 595–608.

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. (2000a). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **56**, 352–359.

Pepe, M. S. (2000b). Receiver operating characteristic methodology. *Journal of the American Statistical Association* **95**, 308–311.

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

Pepe, M. S. and Cai, T. (2004). The analysis of placement values for evaluating discriminatory measures. *Biometrics* **60**, 1099–1107.

Pollard, D. (1990). *Empirical Processes: Theory and Applications.* Hayward, CA: Institute of Mathematical Statistics.

24

Pregibon, D. (1980). Goodness of link tests for generalized linear models. *Applied Statistics* **29**, 15–24.

Rosenfeld, M., Pepe, M., Emerson, J., Longton, G. and FitzSimmons, S. (2001). Effect of different reference equations on the analysis of pulmonary function data in cystic fibrosis. *Pediatirc Pulmonology* **31**, 227–237.

Stute, W. (1997). Nonparametric model checks for regression. *The Annals of Statistics* **25**, 613–641.

Su, J. Q. and Wei, L. J. (1991). A lack-of-fit test for the mean function in a generalized linear model. *Journal of the American Statistical Association* **86**, 420–426.

Swets, J. A. (1986). Indices of discrimination or diagnostic accuracy: Their ROCs and implied models. *Psychological Bulletin* **99**, 100–117.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory.* Academic Press.

Toledano, A. Y. and Gatsonis, C. (1996). Ordinal regression methodology for ROC curves derived from correlated data. *Statistics in Medicine* **15**, 1807–1826.

Tosteson, A. N. and Begg, C. B. (1988). A general regression methodology for roc curve estimation. *Medical Decision Making* **8**, 204–215.

Tsiatis, A. A. (1980). A note on a goodness-of-fit test for the logistic regression model. *Biometrika* **67**, 250–251.

Zheng, Y. and Heagerty, P. (2004). Semiparametric estimation of time-dependent ROC curves for longitudinal marker data. *Biostatistics* **5**, 615–32.

Table 1

Estimated regression coefficients and their standard errors from fitting a location-scale model for
$Y = -FEV_1$ *and* $Y = -\log(FEV_1)$ *in the reference population for the CF registry data*

Covariate	Estimate	SE	Z-statistic	Estimate	SE	Z-statistic
			(a) $Y = -\mathrm{FEV}_1$			
	Mean function			Variance function		
Height	-0.241	0.664	-0.363	-0.095	0.028	-3.419
Age	-1.188	0.208	-5.714	0.002	0.008	0.199
			(b) $Y = \log(-\mathrm{FEV}_1)$			
	Mean function			Variance function		
Height	0.001	0.007	-0.139	-0.130	0.037	-3.495
Age	-0.044	0.017	-2.659	-0.247	0.091	-2.718
Age^2	-0.001	0.001	1.848	0.013	0.004	3.006

26
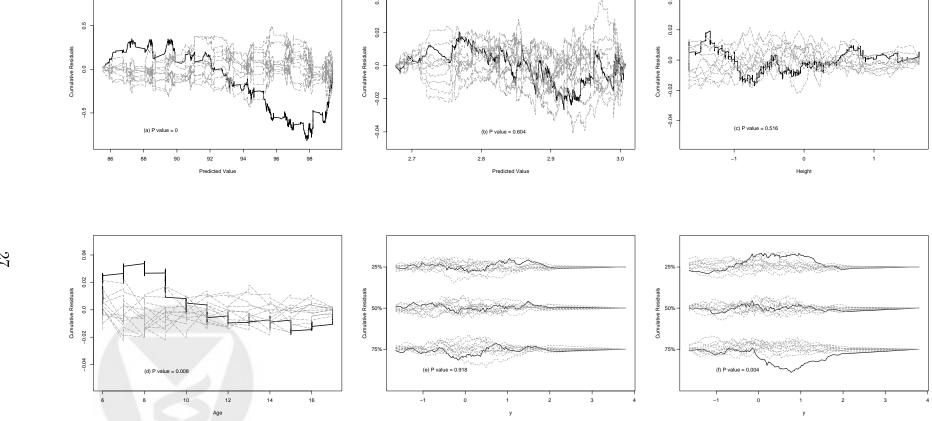
Figure 1. Plots of the cumulative residuals for −FEn controls from the CF registry data. The dark curve shows the observed pattern and the light curves show the 10 simulated realizations. The p-value for the supremum test is obtained with 1000 realizations. (a) $\widehat{M}_{\bar{D}_\mu}(\cdot)$ for examining the linearity in the mean; (b) $\widehat{M}_{\bar{D}_\sigma}(\cdot)$ for examine the variance structure; (c) $\widehat{M}_{\bar{D}_{1_3}}(\cdot)$ for examining the linear height effect; (d) $\widehat{M}_{\bar{D}_{2_3}}(\cdot)$ for examining the linear age effect; (e) $\widehat{M}_{\bar{D}_1}(y, z)$ versus y for height at 25%, 50% and 75% percentile respectively shown from top to bottom; (f) $\widehat{M}_{\bar{D}_2}(y, z)$ versus y for age at 25%, 50% and 75% percentile respectively shown from top to bottom.
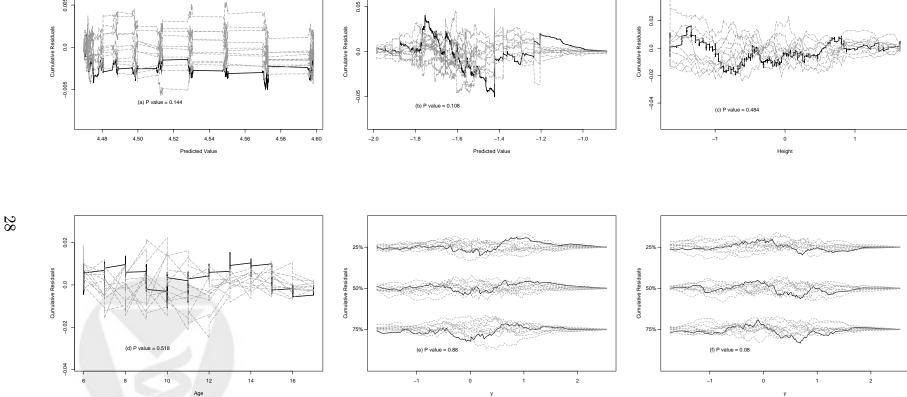
Figure 2. Plots of the cumulative residuals for $log(-FEV_1)$ for controls for the CF registry data. The dark curve shows the observed pattern and the light curves show 10 simulated realizations. The *p*-value for the supremum test is obtained with 1000 realizations. (a) $\widehat{M}_{\bar{D}_\mu}(\cdot)$ for examining the linearity in the mean; (b) $\widehat{M}_{\bar{D}_\sigma}(\cdot)$ for examining the variance structure; (c) $\widehat{M}_{\bar{D}_{1_{\hat{\delta}}}}(\cdot)$ for examining the linear height effect; (d) $\widehat{M}_{\bar{D}_{2_{\hat{\delta}}}}(\cdot)$ for examining the linear age effect; (e) $\widehat{M}_{\bar{D}_1}(y,z)$ versus y for height at 25%, 50% and 75% percentile respectively shown from top to bottom; (f) $\widehat{M}_{\bar{D}_2}(y,z)$ versus y for age at 25%, 50% and 75% percentile respectively shown from top to bottom.
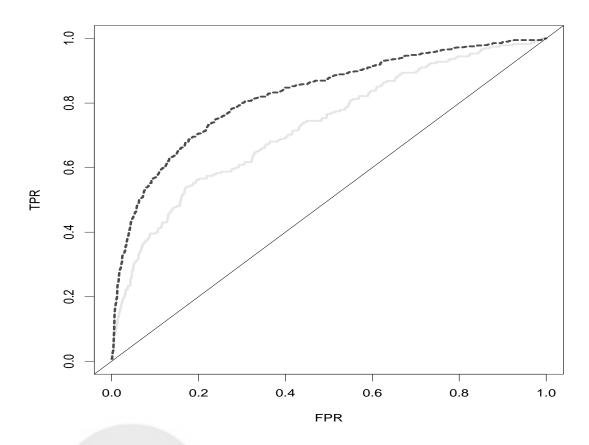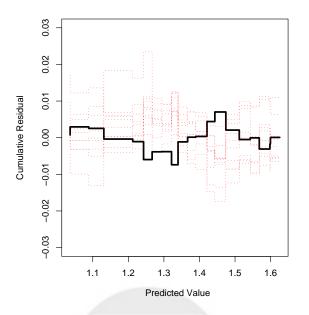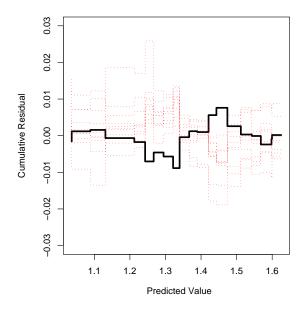
Figure 3. ROC curves estimated by using semiparametric location-scale family FPR and TPR models. The dashed ROC curve is based on the original model for $-\text{FEV}_1$; The solid ROC curve is obtained using a model for $\log(-\text{FEV}_1)$ and with the addition of age^2 in both the location and scale function. The 45 degree line is added for reference.

29

Figure 4. Observed cumulative residual plot based on $\widehat{\mathbf{M}}_{\mathrm{L}}(\cdot)$ (thicker black curve) along with 10 realizations from its null distribution (thinner dashed curves).

(a) Cumulative residual based on $V_{\mathrm{D}ik}$

(b) Cumulative residual based on $V_{\mathrm{D}ik}^2$

30

Figure 5. Estimates (solid curves) of the FPR-specific covariate effects along with their 95% pointwise confidence intervals (dashed curves) and simultaneous confidence bands (dotted curves).

31