Biostatistics (2002), **3**, 3, *pp.* 421–432
Printed in Great Britain

Distribution-free ROC analysis using binary regression techniques

TODD A. ALONZO*
Children's Oncology Group, Keck School of Medicine, University of Southern California, PO Box 60012, Arcadia, CA 91066, USA
talonzo@childrensoncologygroup.org

MARGARET SULLIVAN PEPE
Department of Biostatistics, University of Washington, Seattle, WA 98195, USA
mspepe@u.washington.edu

SUMMARY

Receiver operating characteristic (ROC) regression methodology is used to identify factors that affect the accuracy of medical diagnostic tests. In this paper, we consider a ROC model for which the ROC curve is a parametric function of covariates but distributions of the diagnostic test results are not specified. Covariates can be either common to all subjects or specific to those with disease. We propose a new estimation procedure based on binary indicators defined by the test result for a diseased subject exceeding various specified quantiles of the distribution of test results from non-diseased subjects with the same covariate values. This procedure is conceptually and computationally simplified relative to existing procedures. Simulation study results indicate that the approach has fairly high statistical efficiency. The new ROC regression methodology is used to evaluate childhood measurements of body mass index as a predictive marker of adult obesity.

Keywords: Biomarkers; Classification; Diagnostic tests; Prediction; ROC analysis; Sensitivity; Specificity.

1. INTRODUCTION

Medical diagnostic tests are designed to discriminate between different states of health or medical conditions, e.g. cancer and no cancer. It is well understood that before diagnostic tests are implemented in practice, it is imperative that their accuracy, or ability to discriminate, is studied. Recently, interest has moved beyond determining the basic accuracy of a test. To gain a better understanding of a test, researchers are interested in determining factors that affect its accuracy. In doing so, it is possible to identify populations and settings where a test is more or less accurate, which can be useful in determining how best to use a test. This is accomplished using regression analysis of test accuracy, which is the focus of this paper.

The accuracy of a dichotomous test, a test that yields binary results (i.e. positive or negative), is usually summarized with the true positive rate (TPR) and the false positive rate (FPR). The TPR, also called the sensitivity, is the proportion of diseased subjects correctly detected by the test. On the other hand, FPR or (1-specificity) is defined as the proportion of non-diseased subjects erroneously deemed positive by

*To whom correspondence should be addressed

the test. Receiver operating characteristic (ROC) curves are a well-accepted measure of accuracy for tests that yield ordinal or continuous results. Based on the notion of using a threshold to classify subjects as positive or negative, an ROC curve is a plot of the TPR versus FPR for all possible cutpoints. Thus, it describes the whole range of possible operating characteristics for the test and hence its inherent capacity for distinguishing between diseased and non-diseased states (Zweig and Campbell, 1993).

ROC regression methodology is used to identify factors that affect the discriminating capacity of a non-binary test. We will use ROC regression to investigate the value of monitoring body mass index (BMI) in children to identify children at risk of being obese in adulthood. Childhood BMI is considered as a potential diagnostic marker for adult obesity in our example. Factors potentially influencing its capacity for distinguishing between children deemed to become obese adults include the age of the child when the BMI measurement is made and the sex of the child. Our regression analysis indicates that BMI measured on a young child is not a useful marker although, as one would expect, it becomes a better indicator of risk in older children. Interestingly, BMI may be a better measure of risk in girls than in boys.

A review of the three major existing approaches to ROC regression is provided in Pepe (1998). The first approach uses regression models for the test outcome and infers covariate effects on the corresponding ROC curves (Tosteson and Begg, 1985). This approach has been extended to random effects models (Beam, 1995; Gatsonis, 1995) and Bayesian methods (Peng and Hall, 1996; Hellmich *et al.*, 1998; Ishwaran and Gatsonsis, 2000) when test results are ordinal. The second approach considers regression models for the area under the ROC curve (AUC), a common summary measure of the ROC curve (Thompson and Zucchini, 1989). Finally, a parametric distribution-free (PDF) approach that directly models the ROC curve has been proposed (Pepe, 1997, 2000). In a detailed comparison of the three approaches, Pepe (1998) notes several major advantages to the latter approach, including the facts that it can accommodate multiple test types and continuous covariates and that models can pertain only to restricted portions of the ROC curve that are of interest. Thus, we consider only the direct modelling approach in the remainder of this paper.

The PDF approach assumes a parametric model for the ROC curve but is distribution-free with respect to the distributions of the test results. The key limitation of this approach, as it was originally proposed by Pepe (1997), was that parameter estimation required special programming that made the approach difficult to implement. However, progress towards simplifying the implementation process has been made. Specifically, Pepe (2000) noted that generalized linear model methods for binary data can be used to perform parameter estimation. In this paper, we further simplify the conceptual and computational bases for fitting the PDF model by using as the binary outcome the indicator that a diseased test result is greater than a specified quantile of the distribution of non-diseased observations with the same covariate values. This allows for simpler regression models and for estimation that is less computationally intensive and is easy to implement. Code and data to fit the regression models to the obesity study data are available at the URL provided in Section 6. In addition, we investigate the statistical efficiency of the direct modelling approach for the first time.

This paper is organized as follows. In Section 2, our new formulation of the PDF approach is summarized and relationships with previous methods are noted. Modifications to our formulation are described in Section 3, and results of simulation studies to investigate effects on statistical efficiency are described in Section 4. In Section 5, the methods are applied to data from the obesity study. We end with some recommendations and a discussion.

2. PARAMETRIC DISTRIBUTION-FREE ROC ANALYSIS

Consider the setting where one is interested in determining the effect of a covariate vector X on the accuracy of an ordinal or continuous diagnostic test Y. Let Y_D and $Y_{\bar{D}}$ denote test result random variables

from diseased (D) and non-diseased ($\bar{D}$) populations, respectively. We assume that larger values of Y are more indicative of disease and smaller values are less indicative of disease. It is of interest to determine the effect of covariates, denoted by X and X_D, on the ROC curve, where X represents covariates common to diseased and non-diseased subjects and X_D denotes covariates that are specific to the diseased state. For example, X denotes age and gender in the obesity study and X_D describes a measure of how severe the obesity is in the obese adult.

The ROC curve corresponding to (X, X_D) can be written as $\mathrm{ROC}_{X,X_D}(t) = F_{D,X,X_D}(F^{-1}_{\bar{D},X}(t))$, where $t \in (0,1)$ is the FPR and $F_{D,X,X_D}(c) = P(Y_D \geqslant c|X, X_D)$ and $F_{\bar{D},X}(c) = P(Y_{\bar{D}} \geqslant c|X)$ are survivor functions at threshold c. That is, $\mathrm{ROC}_{X,X_D}(t)$ is the probability that a diseased individual with disease-specific covariates X_D and common covariates X has test results Y_D that are greater than or equal to the tth quantile of the distribution of test results from non-diseased individuals. The general ROC regression model we consider is $\mathrm{ROC}_{X,X_D}(t) = g\big(\sum_{k=1}^{K}\gamma_k h_k(t) + \beta X + \beta_D X_D\big)$. That is, the ROC curve is a function of covariates common to diseased and non-diseased subjects, covariates specific to diseased subjects, and a function $h(\cdot)$ which defines the location and shape of the curve. This approach is referred to as PDF, because the approach specifies a parametric model for the ROC curve but does not assume distributions for the diagnostic test results. The functions $g(\cdot)$ and $h_k(\cdot)$ are chosen so that the ROC curve is monotone increasing on the unit square. In practice, $g(\cdot) = \Phi$ the cumulative normal distribution function, $h_1(t) = 1$, and $h_2(t) = \Phi^{-1}(t)$ are often used. These choices yield the binormal model $\mathrm{ROC}_{X,X_D}(t) = \Phi(\gamma_1 + \gamma_2\Phi^{-1}(t) + \beta X + \beta_D X_D)$ (Metz, 1986). This model specifies that the ROC curves for different values of X and X_D differ by fixed amounts on the probit scale. If $\beta > 0$, then the discrimination between Y_D and $Y_{\bar{D}}$ increases with increasing values of X. Similarly, if $\beta_D > 0$, diseased subjects with larger values of X_D are more distinct from non-diseased subjects than are diseased subjects with smaller values of X_D. A more flexible model could be fit by including an interaction between X or X_D and $\Phi^{-1}(t)$ allowing the effects of (X, X_D) to differ by varying amounts depending on the FPR t.

Pepe (2000) developed a method for fitting this regression model based on the binary indicators $U_{ij} = I[Y_{Di} \geqslant Y_{\bar{D}j}]$ for all pairs of observations $\{(Y_{Di}, Y_{\bar{D}j})i = 1, \ldots, n_D; j = 1, \ldots, n_{\bar{D}}\}$, where n_D and $n_{\bar{D}}$ denote the number of observations for diseased and non-diseased test units, respectively. Here, we consider binary indicators of the form $U_{it} = I[Y_{Di} \geqslant F^{-1}_{\bar{D},X_i}(t)]$ for $t \in T$, where t is a FPR between 0 and 1 and T denotes a fixed finite set of such values. The key observation is

$$\begin{aligned} E[U_{it}] &= P(Y_{Di} \geqslant F^{-1}_{\bar{D},X_i}(t)|X_i, X_{Di}) \\ &= F_{D,X_i,X_{Di}}(F^{-1}_{\bar{D},X_i}(t)) \\ &= \mathrm{ROC}_{X_i,X_{Di}}(t) \\ &= g\left(\sum_{k=1}^{K}\gamma_k h_k(t) + \beta X_i + \beta_D X_{Di}\right), \end{aligned}$$

so that procedures for fitting binary generalized linear models can be applied to $\{U_{it}, i = 1, \ldots, n_D; t \in T\}$ to estimate the model parameters $\{\beta, \beta_D, \gamma_k, k = 1, \ldots, K\}$.

Our algorithm for estimating the model parameters is as follows: (1) specify a set of FPRs, $T = \{t\}$, to consider; (2) estimate $F^{-1}_{\bar{D},X_i}(t)$ for $t \in T$, i.e. calculate the tth covariate-specific quantile of the survivor distribution of the non-diseased test results. This can be accomplished using empirical estimates when applicable or regression quantile methods (e.g. Koenker and Basset, 1978; Heagerty and Pepe, 1999); (3) calculate $U_{it} = I[Y_{Di} \geqslant \hat{F}^{-1}_{\bar{D},X_i}(t)]$ for $i = 1, \ldots, n_D$ and $t \in T$; (4) fit the model $E[U_{it}] = g\big(\sum_{k=1}^{K}\gamma_k h_k(t) + \beta X + \beta_D X_D\big)$ by solving standard estimating equations for fitting a binary generalized

linear model to U_{it} with the link function g^{-1} and covariates $\{h_k(t), X_i, X_{Di};\ k = 1, \ldots, K\}$, i.e.

$$\sum_{i=1}^{n_D} \sum_{t \in T} S_i(\gamma, \beta, \beta_D, t) = \sum_{i=1}^{n_D} \sum_{t \in T} \begin{pmatrix} h(t) \\ X_i \\ X_{Di} \end{pmatrix} w_{\gamma,\beta}(t) \left(U_{it} - g \left\{ \sum_{k=1}^{K} \gamma_k h_k(t) + \beta X_i + \beta_D X_{Di} \right\} \right) = 0, \tag{1}$$

where $w_{\gamma,\beta}(t) = [(\partial/\partial l) g(l)]/g(l)\{1 - g(l)\}$ with $l = \sum_{k=1}^{K} \gamma_k h_k(t) + \beta X + \beta_D X_D$.

Although motivated differently, the estimating equation described by Pepe (2000) for continuous test results is a special case of (1). In particular, with the choice $T = \{\hat{F}_{\bar{D},X_i}(Y_{\bar{D}j}), j = 1, \ldots, n_{\bar{D}}\}$ we show equivalence in the Appendix.

There are several advantages to our formulation. First, we find that a relatively small number of points in T achieve maximal efficiency (see Section 4.3). Thus, the number of binary outcome variables in the new formulation is far fewer than the $n_D \times n_{\bar{D}}$ in the Pepe (2000) formulation. The consequent savings in computational effort can be enormous. Second, the Pepe (2000) formulation requires modelling $E[U_{ij}|X_i, X_{Di}, X_j]$ conditioning on covariates for both the diseased subjects (i) and for the non-diseased subjects (j) in the pair $(Y_{Di}, Y_{\bar{D}j})$ that enter into $U_{ij} = I[Y_{Di} \geqslant Y_{\bar{D}j}]$. There are unfortunate implications for the modelling of continuous covariates. In particular, since the strategy relies on the identity that $E[U_{ij}|X_i, X_{Di}, X_j] = \mathrm{ROC}_{ij}(t_j)$, where $t_j = F_{\bar{D}}(Y_j)$ and $\mathrm{ROC}_{ij}(t_j)$ is the ROC curve that compares diseased subjects with covariates (X_i, X_{Di}) to non-diseased subjects with covariates X_j, it requires modelling how the discrepancy between X_i and X_j affect the ROC curve. Consider as an example an ROC model that includes only $X = \{AGE\}$. Pepe (2000) proposed a model that included $AGE_i - AGE_j$:

$$\mathrm{ROC}_{ij}(t) = g\left(\sum_{k=1}^{K} \gamma_k h_k(t) + \beta_1 AGE_i + \beta_2 (AGE_i - AGE_j) \right).$$

Our new formulation avoids the necessity of modelling discrepancies because it relies on modelling $E[U_{it}|X_i, X_{Di}]$, which conditions only on the covariates for the ith diseased subject, so that the ROC model compares a diseased subject with a non-diseased population at the same covariate value.

A third advantage of the our new approach is that the conceptual framework is simplified relative to previous approaches. Our approach recognizes that the ROC curve compares Y_D with quantiles in an appropriate reference distribution, namely the distribution of non-diseased observations with the same covariate values, $\mathrm{ROC}_{X,X_D}(t) = P(Y_D \geqslant F^{-1}_{\bar{D},X}(t)|X, X_D)$. Based on the identity that $I[Y_D \geqslant F^{-1}_{\bar{D},X}(t)] = I[F_{\bar{D},X}(Y_D) \geqslant t]$, another interpretation of our algorithm is that it calculates the placement values for Y_D in the reference distribution $F_{\bar{D},X}$ and compares them with $t \in (0, 1)$. Thus, the ROC curve at t is recognized as the proportion of such placement values exceeding t.

3. A SYMMETRIZED FITTING PROCEDURE

Our new approach compares the observations for diseased subjects with their reference distributions derived from the non-diseased set. One could, alternatively, compare the observations for non-diseased subjects with reference distributions derived from the disease set. Indeed, one could do both and perhaps arrive at a composite procedure that is more efficient than either is alone. We considered this avenue for a simple setting.

The classic binormal ROC model without covariates is

$$\mathrm{ROC}(t) = \Phi(\alpha + \beta \Phi^{-1}(t)), \tag{2}$$

where t is the FPR. Alternatively, we could consider the ROC curve as a function of the TPR instead of the FPR. This would yield the following model

$$\text{ROC}^{\star}(s) = \Phi(\alpha^{\star} + \beta^{\star}\Phi^{-1}(s)), \tag{3}$$

where s is the TPR, $\alpha^{\star} = -\alpha\beta^{-1}$, and $\beta^{\star} = \beta^{-1}$. Thus, (α, β) in (2) can also be estimated by fitting (3). Specifically, $(\alpha^{\star}, \beta^{\star})$ can be estimated by interchanging the labels for D and $\bar{D}$ in the algorithm outlined in Section 2 and calculating the estimates $\hat{\alpha^{\star}}$ and $\hat{\beta}^{\star}$ and the corresponding estimates of (α, β) as $\hat{\alpha} = -\hat{\alpha^{\star}}\hat{\beta^{\star}}^{-1}$ and $\hat{\beta} = \hat{\beta^{\star}}^{-1}$. We refer to this as the interchanged approach since the roles of D and $\bar{D}$ have been interchanged.

Let $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$ correspond to (α, β) estimated using the models (2) and (3), respectively. Now that two methods have been identified for estimating (α, β), the question arises as to how to optimally combine the estimates. We calculate a new estimate $(\hat{\alpha}_S, \hat{\beta}_S)$ as a weighted average of the two estimates. The minimum variance weighted averages of $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$ are $\hat{\alpha}_S = \frac{(\sigma_1^2+\sigma_{12}^2)\hat{\alpha}_1+(\sigma_2^2+\sigma_{12}^2)\hat{\alpha}_2}{\sigma_1^2+\sigma_2^2+2\sigma_{12}^2}$ and $\hat{\beta}_S = \frac{(\tau_1^2+\tau_{12}^2)\hat{\beta}_1+(\tau_2^2+\tau_{12}^2)\hat{\beta}_2}{\tau_1^2+\tau_2^2+2\tau_{12}^2}$, where $\text{Cov}(\hat{\alpha}_1, \hat{\alpha}_2) = \begin{pmatrix} \sigma_1^2 & \sigma_{12}^2 \\ \sigma_{12}^2 & \sigma_2^2 \end{pmatrix}$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2) = \begin{pmatrix} \tau_1^2 & \tau_{12}^2 \\ \tau_{12}^2 & \tau_2^2 \end{pmatrix}$. In practice, (α_1, β_1) and (α_2, β_2) can be estimated simultaneously, and the bootstrap can be used to obtain estimates of $\text{Cov}(\hat{\alpha}_1, \hat{\alpha}_2)$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$. It has been our experience that 500 bootstrap samples have been adequate.

This estimation procedure has the advantage that it is symmetric in the use of Y_D and $Y_{\bar{D}}$, i.e. the arbitrary labeling 'diseased' versus 'non-diseased' does not matter. Hence, we refer to this method as the symmetrized approach.

4. Simulations

We performed simulation studies to investigate questions about statistical efficiency. For this purpose, we considered models without covariates. Continuous test results from a non-diseased population, $Y_{\bar{D}}$, were generated using a standard normal random variable. Y_D were generated from a normal random variable with mean $\alpha\beta^{-1}$ and variance β^{-2}. This formulation yields the classic binormal ROC curve (2).

We simulated two different scenarios: (i) $\alpha = 0.75$, $\beta = 0.90$ (AUC = 0.711) and (ii) $\alpha = 1.5$, $\beta = 0.85$ (AUC = 0.874). The corresponding ROC curves for these two scenarios are displayed in Figure 1 and are representative of curves that may be encountered in practice. An equal number of diseased and non-diseased observations were used in each simulation.

4.1 *Comparison with symmetrized approach*

First, we evaluated the gain in statistical efficiency from estimating the binormal parameters in model (2) by using the symmetrized procedure of Section 3. Five-hundred bootstrap samples were used in each simulation to estimate the covariances $\text{Cov}(\hat{\alpha}_1, \hat{\alpha}_2)$ and $\text{Cov}(\hat{\beta}_1, \hat{\beta}_2)$ necessary to combine the estimates $(\hat{\alpha}_1, \hat{\beta}_1)$ and $(\hat{\alpha}_2, \hat{\beta}_2)$. The jackknife-after-bootstrap (Efron, 1992) suggested 500 bootstrap samples was adequate. Table 1 summarizes the results of 500 simulations with 50 diseased and 50 non-diseased subjects. The maximal set $T = \{j/n_{\bar{D}}; j = 1, \ldots, n_{\bar{D}} - 1\}$, i.e. cutpoints at every observed non-diseased observation, was used with both the original and interchanged estimators.

The results indicate that the symmetrized procedure has statistical properties that are very similar to the original PDF procedure described in Section 2. The standard deviations and mean-squared errors (MSEs) for $(\hat{\alpha}_S, \hat{\beta}_S)$ and $(\hat{\alpha}_{\text{PDF}}, \hat{\beta}_{\text{PDF}})$ are virtually identical. This suggests that there is no additional information gleaned regarding the ROC curve from the non-diseased observations in the interchanged approach that is not already contained in the estimating equation (1).

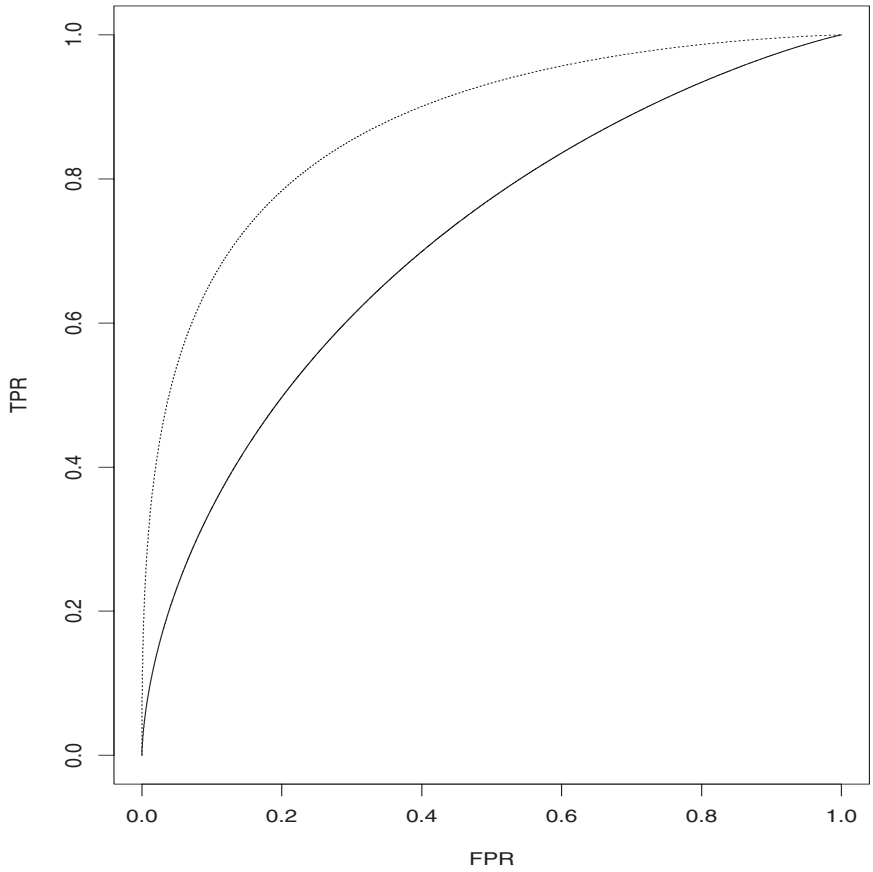

Fig. 1. Binormal ROC curves considered in the simulation studies. Scenario (i) $\alpha = 0.75$, $\beta = 0.90$ (solid curve) and scenario (ii) $\alpha = 1.5$, $\beta = 0.85$ (broken curve).

Table 1. *A summary of the estimates of the binormal ROC model parameters based on* $n_D = n_{\bar{D}} = 50$ *observations in* 500 *simulated datasets. Results are shown for the PDF estimator from* (1)*, interchanged estimator* (*I*)*, and the symmetrized estimator* (*S*)*. Summaries shown are mean, standard deviation* (*SD*)*, and MSE*

	$\alpha = 0.75, \beta = 0.9$			$\alpha = 1.5, \beta = 0.85$		
Estimator	Mean	SD	MSE	Mean	SD	MSE
$\widehat{\alpha}_{\text{PDF}}$	0.735	0.222	0.050	1.529	0.298	0.090
$\widehat{\alpha}_I$	0.814	0.231	0.057	1.610	0.292	0.097
$\widehat{\alpha}_S$	0.776	0.227	0.052	1.571	0.289	0.089
$\widehat{\beta}_{\text{PDF}}$	0.926	0.167	0.029	0.913	0.210	0.048
$\widehat{\beta}_I$	0.936	0.168	0.029	0.889	0.196	0.040
$\widehat{\beta}_S$	0.932	0.165	0.028	0.901	0.196	0.041

Table 2. *A summary of the estimates of parameters of the binormal ROC model in* 3000 *simulations with* $n_D = n_{\bar{D}} = 50$ *using the PDF and ML estimators. Continuous test results were categorized into five ordinal categories*

	$\alpha = 0.75, \beta = 0.9$			$\alpha = 1.5, \beta = 0.85$		
Estimator	Mean	SD	MSE	Mean	SD	MSE
$\widehat{\alpha}_{\text{PDF}}$	0.754	0.233	0.054	1.561	0.361	0.134
$\widehat{\alpha}_{\text{ML}}$	0.778	0.234	0.055	1.590	0.372	0.146
$\widehat{\beta}_{\text{PDF}}$	0.902	0.204	0.041	0.898	0.307	0.097
$\widehat{\beta}_{\text{ML}}$	0.921	0.211	0.045	0.919	0.317	0.105

4.2 *Comparison with maximum likelihood*

Next, we further investigated the efficiency of the PDF approach. We chose to do this for ordinal test results because in this setting maximum likelihood (ML) estimators of the binormal model parameters are established and fully efficient (Dorfman and Alf, 1969). To simulate ordinal test results, we categorized the continuous test results Y_D and $Y_{\bar{D}}$ into five ordinal categories using the four cutpoints $\Phi^{-1}(0.3)$, $\Phi^{-1}(0.5)$, $\Phi^{-1}(0.7)$, and $\Phi^{-1}(0.9)$.

Table 2 summarizes estimates of the binormal model parameters in 3000 simulations of studies with $n_D = n_{\bar{D}} = 50$. The maximal FPR set T was used with the PDF approach, i.e. T consists of the four observed FPRs corresponding to thresholds at each of the ordinal categories. The PDF and ML approaches have similar efficiency and MSE. This suggests that the PDF estimator is in fact reasonably efficient. Moreover, the result in Section 4.1, that symmetrizing the approach did not substantially improve the efficiency, is no longer surprising, because we see that the original distribution-free approach is already efficient.

4.3 *The choice of the false positive rate set, T*

Our PDF approach requires specifying T, the set of FPRs. Next, we investigated how the choice of T affects the efficiency of our estimator. The maximal set of FPRs is $T = \{j/n_{\bar{D}}; j = 1, \ldots, n_{\bar{D}} - 1\}$ corresponding to cutpoints at every observed non-diseased observation. This setting yields the estimator described in Pepe (2000). We considered sets of n_T equally spaced values in (0, 1) for various values of n_T. Data were simulated for $n_D = n_{\bar{D}} = 500$ subjects from the binormal model (2).

Table 3 displays the ratio of the variance of the parameter estimates obtained with n_T in the range 4 to 50 relative to the maximal $n_T = 499$. We find good efficiency in estimating the intercept parameter, α, even for small n_T. For example, when $T = \{1/11, 2/11, \ldots, 10/11\}$ (i.e. $n_T = 10$), the approach that uses 50 times fewer points is only 2 and 3% less efficient than the maximal T in scenarios (i) and (ii), respectively. Although a larger set is required to achieve good efficiency in the estimation of the intercept parameter, β, relative efficiencies of 0.95 and 0.90 were obtained even with $n_T = 50$, which is one-tenth the size of the maximal T.

Models for the ROC curve in a restricted range of FPRs between 0 and 0.2 were also considered (results not presented). Our approach can fit such models by restricting T to (0, 0.2). Efficiencies of at least 80 and 90% relative to the maximal $n_T = 99$ were observed for $n_T = 10$ and $n_T = 15$ equally spaced points, respectively. Thus, a relatively small number of FPRs in T achieve near-maximal efficiency for model-based estimates of ROC curves in restricted and unrestricted ranges of FPRs.

Table 3. *Efficiency of binormal model parameter estimates relative to maximal set of FPRs. The number of equally spaced FPRs, n_T, is varied*

n_T	$\alpha = 0.75$	$\beta = 0.9$	$\alpha = 1.5$	$\beta = 0.85$
4	0.93	0.47	0.88	0.39
7	0.98	0.68	0.96	0.55
10	0.98	0.77	0.97	0.66
15	0.99	0.86	0.98	0.74
20	0.99	0.87	0.97	0.79
40	1.00	0.93	0.99	0.87
50	1.00	0.95	0.99	0.90
499	1.00	1.00	1.00	1.00

5. APPLICATION TO OBESITY DATA

A primary goal of the Childhood Predictors of Adult Obesity Study (CPAO) was to determine how well childhood obesity could predict the likelihood of adult obesity. CPAO was a retrospective observational study of subjects born at a health maintenance organization between 1965 and 1971 and still members when they were adults (age $\geqslant$ 21 years). Detailed eligibility criteria are given in Whitaker *et al.* (1997).

Data were available on 823 adults, 305 (37.1%) male and 518 (62.9%) female. The average BMI, defined as the weight in kilograms divided by the square of height in meters, between 21 and 29 years of age was calculated for each adult in the study. Based on established guidelines, an adult was then classified as obese if the average BMI exceeded 27.3 and 27.8 kg m^{-2} for females and males, respectively (NIH Consensus Statement, 1985). The prevalence of adult obesity in this cohort is $133/823 = 16.2\%$.

6091 childhood BMI measurements between ages 3 and 18 were available on the study subjects. A median of 7.2 childhood BMI measurements were collected per person (range 1–26). Childhood BMI values were standardized for age and gender by conversion to a z-score using age and gender-specific means and standard deviations from a reference population (Frisancho, 1990).

Using the notation introduced in Section 2, D corresponds to the indicator of adult obesity and Y is the childhood BMI z-score. Each person contributes several observations to the analysis, one at each time that childhood BMI was measured. Questions of interest are to determine at different ages how well childhood BMI discriminates people that become obese in adulthood from those that do not and if the measure is better in girls than in boys. We also asked how well BMI would distinguish persons who become severely obese from those that remain in normal limits of BMI throughout adulthood. Therefore, the covariates of interest, X, are the age corresponding to the BMI z-score (*AGE*), gender (*GENDER*), and X_D is adult BMI z-score (*aBMIz*).

The following ROC regression model was fitted:

$$\mathrm{ROC}_{X,X_D}(t) = \Phi(\gamma_1 + \gamma_2\Phi^{-1}(t) + \beta X + \theta X\Phi^{-1}(t) + \beta_D X_D + \theta_D X_D\Phi^{-1}(t)).$$

This model is very flexible, in that it allows the effects of (X, X_D) on the ROC curves to differ by varying amounts depending on the FPR t. Parameter estimation suggested that interactions between covariates and $\Phi^{-1}(t)$ were not significant and, thus, the following reduced model was more appropriate:

$$\mathrm{ROC}_{X,X_D}(t) = \Phi(\gamma_1 + \gamma_2\Phi^{-1}(t) + \beta X + \beta_D X_D). \quad (4)$$

This model implies that ROC curves for different values of X differ by fixed amounts on the probit scale and therefore cannot cross.

Table 4. *Results of ROC regression analysis applied to the CPAO study*

Variable	Coefficient	Standard error	p-value
Intercept	0.210	0.225	0.348
AGE (years)	0.080	0.014	<0.0001
GENDER (female = 0, male = 1)	−0.313	0.185	0.090
aBMIz (z-score)	0.285	0.084	0.001
$\Phi^{-1}(t)$	1.140	0.069	<0.0001

Of the 6091 records, 5165 and 926 correspond to non-obese and obese adults, respectively. Thus, $n_{\bar{D}} = 5165$ and $n_D = 926$. Application of the fitting procedure described in Pepe (2000) is not practical with these data because it requires probit regression to be performed on more than 4 million observations (i.e. $n_D \times n_{\bar{D}}$). Instead, we applied our computationally less intensive modified approach described in Section 2. Since simulation results suggested that the symmetrized approach did not gain much efficiency, we chose not to symmetrize the estimation procedure.

Using the algorithm outlined in Section 2, we first specified the set of FPRs to be used. Based on the simulation results summarized in Section 4.3, 50 equally spaced FPRs, $T = \{1/51, 2/51, \ldots, 50/51\}$, were used. Next, quantiles of the survivor distribution of $Y_{\bar{D}}$, childhood BMI z-scores for the non-obese adults, were estimated as a function of *AGE* and *GENDER* for $t \in T$ using regression quantile methods (Koenker and Basset, 1978). Then $U_{it} = I[Y_{Di} \geqslant F^{-1}_{\bar{D},X_i}(t)]$ was calculated for $i = 1, \ldots, 926$ and $t \in T$. A probit regression model was then fitted to U_{it} with covariates $\Phi^{-1}(t)$, $X = (AGE, GENDER)$, and $X_D = aBMIz$. The results are summarized in Table 4. Standard errors were estimated using 500 bootstrap samples, with the unit for resampling being the cluster of data for the study subject. The jackknife-after-bootstrap (Efron, 1992) suggested 500 bootstrap samples was adequate.

The negative coefficient for *GENDER* suggests that childhood BMI is a more accurate predictor of adult obesity in females than in males, although this result is not conclusive ($p = 0.090$). Since the coefficient estimate for *AGE* is positive, it implies that the older the child is at the time of the BMI measurement, the better BMI is for discriminating those that will be obese as adults. This is to be expected because the older the child, the closer the child is to adulthood. The positive coefficient for *aBMIz* implies that it is easier to distinguish severely obese adults from non-obese adults on the basis of their childhood values than to distinguish moderately obese adults from non-obese adults.

ROC curves for several different profiles of children in the study are displayed in Figure 2. The farther the ROC curve is from the 45° line, the better childhood BMI is in predicting adult obesity. Figure 2 suggests that BMI measurements obtained on 3-year-olds will not distinguish between children who become borderline obese ($aBMIz = 1.04$) and children who have normal BMI in adulthood. The corresponding estimates of AUC are 0.637 and 0.707 for males and females, respectively. Conversely, BMI measurements collected at age 13 for borderline obese adult males (AUC = 0.792) and females (AUC = 0.846) are much better markers of risk. Moreover, children deemed to become severely obese ($aBMIz = 2.5$) are already at age 13 very different from their peers who will remain within normal BMI limits in adulthood (AUC = 0.862 for males, AUC = 0.902 for females). The point on the ROC curve that yields a TPR of 0.8 for 13-year-old girls that become severely obese adults corresponds to the BMI z-score threshold of 0.33 and a FPR of 0.17 in 13-year-old girls that do not become obese adults. Thus, a dietary intervention on 13-year-old girls whose BMI z-score exceeds 0.33 would reach 80% of the girls who would be severely obese in adulthood and would unnecessarily subject to intervention 17% of the girls that would reach adulthood in a non-obese state.

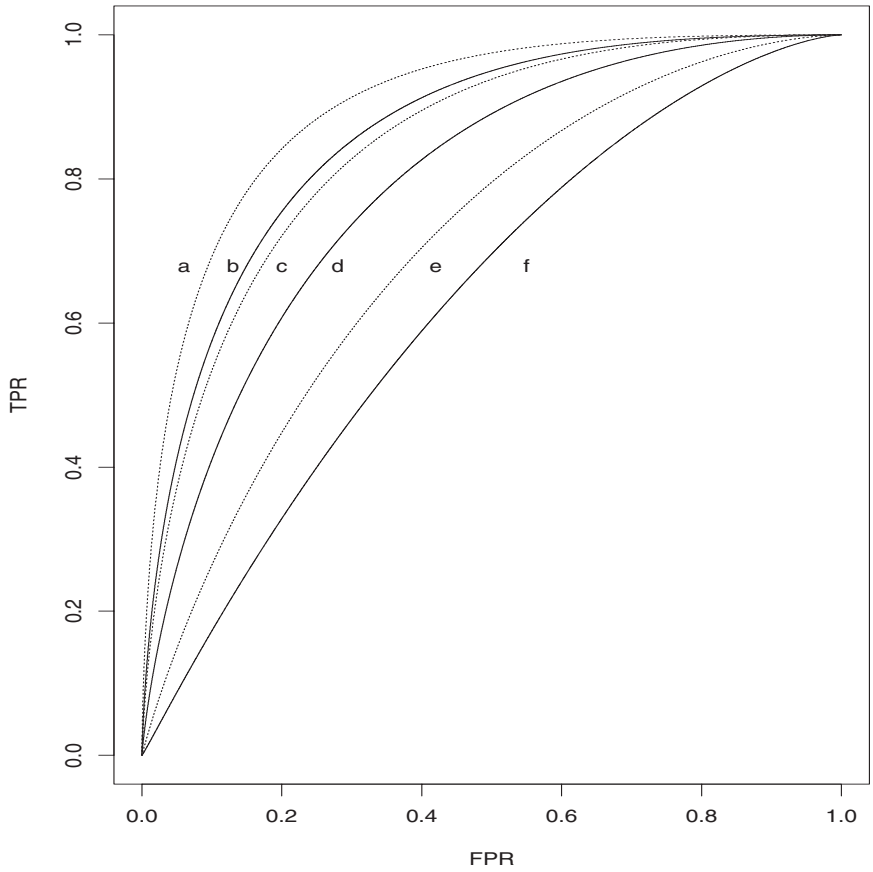

Fig. 2. ROC curves for childhood BMI z-scores as a marker of risk for obesity in adulthood. Males (solid curve) and females (broken curve). 13-year-olds who become severely obese adults (a, b); 13-year-olds who are borderline obese adults (c, d); 3-year-olds who are borderline obese adults (e, f).

6. DISCUSSION

We have previously used ROC regression to compare two ROC curves for continuous biomarkers (Pepe, 2000), for comparing two imaging methods in a multi-reader comparative radiology study (Pepe, 1997), and for evaluating multiple covariates affecting an audiology diagnostic test (Pepe, 1998). In Section 5, we considered a more elaborate example involving a longitudinal study to evaluate childhood BMI as a potential predictive marker for adult obesity. This sort of application is important as the search for early biomarkers of diseases becomes a popular area of research and longitudinal data sets become available to investigate their value.

ROC regression analysis was used in Section 5 to determine how well childhood obesity can predict adult obesity. This is a retrospective analysis, in that it corresponds to TPR and FPR which are accuracy measures that condition on disease status. Alternatively, a prospective analysis approach using predictive values which condition instead on childhood BMI z-value could be used. Pepe *et al.* (1999) performed such an analysis on these data. Either approach is valid. In evaluating diagnostic tests, TPR and FPR are often preferred because they are independent of prevalence and of factors influencing prevalence alone.

They are also more descriptive of how the test reflects disease status.

The ROC regression model proposed in this paper allows one to simultaneously compare multiple subsets of diseased subjects with a common non-diseased population. A disease-specific covariate that we denote by X_D can indicate the subpopulation, which could refer to disease subtype or characteristics of disease. In this sense, the formulation is applicable to settings where there are more than two disease states. In Section 5, for example, ROC curves were constructed for 'borderline obese' and 'severely obese' patients.

In summary, we propose using a modified parametric distribution-free ROC estimator which is conceptually easy and is simple to implement with existing software packages. Code and data for the obesity example can be found at `http://www-rcf.usc.edu/~talonzo/dfROC.html`. Results from simulation studies indicated that this approach is reasonably efficient and, thus, there is no need to symmetrize the estimation. Furthermore, simulation studies indicate that fairly high efficiency can be attained using a small number of equally spaced FPRs to fit the regression models. Future research in this area might include the development of an algorithm for optimally choosing an efficient set of FPRs. Criteria for selecting this set could, for example, depend on the distribution of data in particular regions of the ROC curve.

ACKNOWLEDGEMENTS

This research is partially funded by NIH/NCI 5U10 CA13539 and NIH R01 GM54438.

APPENDIX

Consider the estimating equation (5) of Pepe (2000) which in our notation is written as

$$\sum_{i=1}^{n_D}\sum_{j=1}^{n_{\bar{D}}}\begin{pmatrix} h(\hat{t}_j) \\ X_{ij} \end{pmatrix} w_{\gamma,\beta}(\hat{t}_j)\left(U_{ij} - g\left\{\sum_{k=1}^{K}\gamma_k h_k(\hat{t}_j) + \beta X_{ij}\right\}\right) = 0,$$

where $U_{ij} = I[Y_{Di} \geqslant Y_{\bar{D}j}]$ and $\hat{t}_j = \hat{F}_{\bar{D},X_j}(Y_{\bar{D}j})$. The X_{ij} are covariables defined by covariates for the ith diseased and jth non-diseased paired observations (X_i, X_j). Suppose that $X_i = X_j$ or that X_j is the null set. Then we can write $X_{ij} = X_i$ and $\hat{t}_j = \hat{F}_{\bar{D},X_j}(Y_{\bar{D}j})$. The above expression becomes that in (1) if we define $T = \{\hat{t}_j,\ j = 1, \ldots, n_{\bar{D}}\}$.

REFERENCES

BEAM, C. A. (1995). Random-effects models in the receiver operating characteristic curve-based assessment of the effectiveness of diagnostic imaging technology: concepts, approaches, and issues. *Academic Radiology* **2**, S4–S13.

DORFMAN, D. D. AND ALF, E. (1969). Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals: rating data. *Journal of Mathematical Psychology* **6**, 487–496.

EFRON, B. (1992). Jacknife-after-bootstrap standard errors and influence functions. *Journal of the Royal Statistical Society,* Series B **54**, 83–127.

FRISANCHO, A. R. (1990). *Anthropometric Standards for the Assessment of Growth and Nutritional Status*. Ann Arbor: University of Michigan Press.

GATSONIS, C. A. (1995). Random-effects models for diagnostic accuracy data. *Academic Radiology* **2**, S14–S21.

HEAGERTY, P. J. AND PEPE, M. S. (1999). Semiparametric estimation of regression quantiles with application to standardizing weight for height and age in children. *Applied Statistics* **48**, 533–551.

HELLMICH, M., ABRAMS, K. R., JONES, D. R. AND LAMBERT, P. C. (1998). A Bayesian approach to a general regression model for ROC curves. *Medical Decision Making* **18**, 436–443.

ISHWARAN, H. AND GATSONSIS, C. A. (2000). A general class of hierarchical ordinal regression models with applications to correlated ROC analysis. *Canadian Journal of Statistics* **28**, 731–750.

KOENKER, R. AND BASSET, G. (1978). Regression quantiles. *Econometrica* **46**, 33–50.

METZ, C. E. (1986). ROC methodology in radiologic imaging. *Investigative Radiology* **21**, 720–733.

NIH CONSENSUS STATEMENT (1985). National Institutes of Health consensus development panel on the health implications of obesity. *Annals of Internal Medicine* **103**, 1073–1077.

PENG, F. AND HALL, W. J. (1996). Bayesian analysis of ROC curves using Markov-chain Monte Carlo methods. *Medical Decision Making* **16**, 404–411.

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

PEPE, M. S., HEAGERTY, P. AND WHITAKER, R. (1999). Prediction using partly conditional time-varying coefficients regression models. *Biometrics* **55**, 944–950.

THOMPSON, M. L. AND ZUCCHINI, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine* **8**, 1277–1290.

TOSTESON, A. AND BEGG, C. B. (1985). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**, 204–215.

WHITAKER, R. C., WRIGHT, J. A., PEPE, M. S., SEIDEL, K. D. AND DIETZ, W. H. (1997). Predicting obesity in young adulthood from childhood and parent obesity. *New England Journal of Medicine* **337**, 869–873.

ZWEIG, M. H. AND CAMPBELL, G. (1993). Receiver-operating characteristic (ROC) plots: a fundamental evaluation tool in clinical medicine. *Clinical Chemistry* **39**, 561–577.

[*Received April 26, 2001*; *revised August 24, 2001*; *accepted for publication September 19, 2001*]