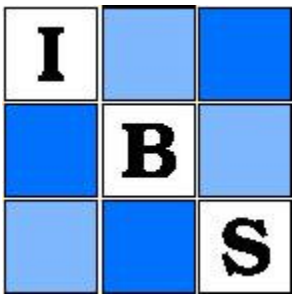

REFERENCES

BIOMETRICS 60, 528–535
June 2004

The Analysis of Placement Values for Evaluating Discriminatory Measures

Margaret Sullivan Pepe

Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
email: mspepe@u.washington.edu

and

Tianxi Cai

Department of Biostatistics, Harvard School of Public Health, Boston, Massachusetts 02115, U.S.A.

SUMMARY. The idea of using measurements such as biomarkers, clinical data, or molecular biology assays for classification and prediction is popular in modern medicine. The scientific evaluation of such measures includes assessing the accuracy with which they predict the outcome of interest. Receiver operating characteristic curves are commonly used for evaluating the accuracy of diagnostic tests. They can be applied more broadly, indeed to any problem involving classification to two states or populations ($D = 0$ or 1). We show that the ROC curve can be interpreted as a cumulative distribution function for the discriminatory measure Y in the affected population ($D = 1$) after Y has been standardized to the distribution in the reference population ($D = 0$). The standardized values are called placement values. If the placement values have a uniform(0, 1) distribution, then Y is not discriminatory, because its distribution in the affected population is the same as that in the reference population. The degree to which the distribution of the standardized measure differs from uniform(0, 1) is a natural way to characterize the discriminatory capacity of Y and provides a nontraditional interpretation for the ROC curve. Statistical methods for making inference about distribution functions therefore motivate new approaches to making inference about ROC curves. We demonstrate this by considering the ROC-GLM regression model and observing that it is equivalent to a regression model for the distribution of placement values. The likelihood of the placement values provides a new approach to ROC parameter estimation that appears to be more efficient than previously proposed methods. The method is applied to evaluate a pulmonary function measure in cystic fibrosis patients as a predictor of future occurrence of severe acute pulmonary infection requiring hospitalization. Finally, we note the relationship between regression models for the mean placement value and recently proposed models for the area under the ROC curve which is the classic summary index of discrimination.

KEY WORDS: Classification; Prediction; Receiver operating characteristic curves; Sensitivity; Specificity.

1. Introduction

This article is about evaluating measures that promise to discriminate between defined populations. For example, diagnostic tests seek to distinguish subjects with a condition from those without. Screening tests seek to distinguish subjects who will develop a serious condition from those who will not. New technologies, including genomic and proteomic measurements, are intended for several classification purposes. Examples include predicting the prognosis of patients with disease (van't Veer et al., 2002) and predicting their response to therapy (Elmer-Dewitt et al., 2001).

A statistical methodology to quantify the discriminatory capacity of such measures is well developed in the context of evaluating diagnostic tests (Zhou, Obuchowski, and McClish, 2002; Pepe, 2003). ROC methodology can also be adopted more broadly (Swets, 1988). In this article, we provide a view of ROC analysis that is not widely known. This nontraditional viewpoint may render ROC analysis more accessible to some researchers and clarify its potential for broader applicability. The key idea is that the ROC curve is in fact the probability distribution of placement values (Hanley and Haijian-Tilaki, 1997), where placement values are defined as standardized versions of the raw measurements. This leads us to propose that standard statistical methods for inference about probability distributions can be applied to the placement values to make inference about ROC curves. New methodology for ROC analysis ensues.

The methods are used to evaluate FEV_1, a measure of lung function, for predicting acute pulmonary exacerbations in children with cystic fibrosis. We find that it discriminates better in patients who are less healthy.

2. Placement Values and the ROC Curve

Let Y be the continuous valued measure that is sought to discriminate between two populations. The indicator variable D denotes the population, $D = 0$ or 1, and we assume larger values of Y are associated with $D = 1$. For example, the populations could be those diseased ($D = 1$) versus nondiseased ($D = 0$), in diagnostic testing or those who die of their illness versus those who do not, in prognostic research. We choose one of the populations, $D = 0$, as the *reference* and use subscript $\bar{D}$ to index-related quantities. Let $F_{\bar{D}}(y)$ denote the cumulative distribution function (cdf) of Y in the reference population. The *placement value* of Y is

$$U = 1 - F_{\bar{D}}(Y).$$

It is the proportion of the reference population with values larger than Y and is simply a transformation of Y that standardizes Y to the distribution in the reference population. Interestingly, this standardization is commonly used already in some areas of medicine. For example, a child's weight is usually reported as the percentile to which it corresponds in a healthy population (Hamill et al., 1977). If the child's weight is at the 90th percentile, then the equivalent placement value is 10%.

We call the population $D = 1$ the *affected* population, and we use the corresponding subscript D. The distribution of placement values in the reference population, $U_{\bar{D}} = 1 - F_{\bar{D}}(Y_{\bar{D}})$, is uniform(0, 1) by definition. On the other hand, the distribution of placement values in the affected population, $U_D = 1 - F_{\bar{D}}(Y_D)$, quantifies the separation between the populations. If the populations are highly separated, the placement of most affected subjects is at the upper tail of the reference distribution, so that most will have small placement values. If the populations overlap substantially, larger placement values will be more common and the cdf of U_D will not rise so steeply. In Figure 1 we illustrate several scenarios.

The ROC curve is typically used to summarize the discriminatory capacity of a diagnostic test (Swets and Pickett, 1982). It is a plot of the true versus false positive rates for classification rules based on Y, i.e., the plot of $1 - F_D(y) = P(Y > y \mid D = 1)$ versus $1 - F_{\bar{D}}(y) = P(Y > y \mid D = 0)$ for all threshold values $y \in (-\infty, \infty)$ that could be used to define test positivity. Setting $u = 1 - F_{\bar{D}}(y)$ we see that this is the function

$$\mathrm{ROC}(u) = 1 - F_D\big(F_{\bar{D}}^{-1}(1-u)\big), \quad u \in (0,1).$$

Observe that the cdf for the placement values, U_D, is

$$\begin{aligned} P(U_D \le u) &= P(1 - F_{\bar{D}}(Y) \le u \mid D = 1) \\ &= P\big(F_{\bar{D}}^{-1}(1-u) \le Y \mid D = 1\big) \\ &= 1 - F_D\big(F_{\bar{D}}^{-1}(1-u)\big) \\ &= \mathrm{ROC}(u). \end{aligned}$$

Therefore, the ROC curve can also be interpreted as the distribution of Y in the affected population when Y is standardized relative to the reference population in a natural way (Figure 1).

When covariates $\mathbf{Z}$ affect the distribution of Y in the reference population, let $F_{\bar{D},\mathbf{Z}}$ denote the covariate specific reference distribution function. Then the placement value for a subject with covariates $\mathbf{Z}$ is defined as

$$U = 1 - F_{\bar{D},\mathbf{Z}}(Y).$$

The distribution of placement values may or may not depend on covariates. If the distribution does not depend on covariates this implies that the discrimination between reference and affected populations (with the same covariate values) does not vary with covariates.

However, in practice, a measure might discriminate better in certain settings than in others. For example, covariates that denote variations on the assay technique or protocol for ascertaining Y could affect the discriminatory capacity of Y. See Pepe (2003) for more examples. Regression models for U_D can be used to quantify such covariate effects on discrimination. One such model is

$$H_{\boldsymbol{\alpha}}(U_D) = -\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + \varepsilon, \tag{1}$$

where ε has a specified distribution function g, and $H_{\boldsymbol{\alpha}}$ is a parametric increasing function with intercept. It is easy to show that this model is equivalent to the class of ROC regression models proposed by Pepe (1997) that are of the form

$$\mathrm{ROC}_{\mathbf{Z}}(u) = g\big\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z} + H_{\boldsymbol{\alpha}}(u)\big\}. \tag{2}$$

By writing (2) as a model for the distribution of placement values in (1), we have a new conceptual framework for the evaluation of covariates on discrimination. In addition, we will show in Section 3 that it gives rise to new techniques for making inference about ROC regression parameters.

The most common summary index of the ROC curve is the area under the curve (AUC). Because the expected value of a random variable is the area under its survival (1-cumulative distribution) function, we see that the mean of the placement value distribution is therefore related to the AUC:

$$\mathrm{AUC} = E(1 - U_D).$$

Traditionally, ROC curves for two measures Y_1 and Y_2 are compared by calculating the difference between their estimated AUCs. Another interpretation is that one first standardizes each of the measures to their respective reference distributions by calculating placement values, and then one compares the means of the standardized values. This latter interpretation fits well with traditional mainstream statistical methodology. This interpretation of the AUC is not new. A paper by Hanley and Hajian-Tilaki (1997) provides a nice discussion. The point we wish to stress here is that ROC analysis in general can be viewed as the analysis of measures standardized as placement values. This has not been fully exploited in the statistical literature. There are many avenues to pursue. In Section 3 of this article, we develop a new method based on the likelihood of the placement values to make inference about ROC regression models. In Section 6, we use the mean placement value interpretation of the AUC to develop a new procedure for making inference about covariate effects on the AUC.

3. The Pseudo-Likelihood Function

Recall the ROC regression model in equation (2) or equivalently the placement value model in equation (1). To be precise with notation for covariates, let $\mathbf{Z}$ denote covariates that affect Y in the reference distribution and write $F_{\bar{D},\mathbf{Z}}$ for the

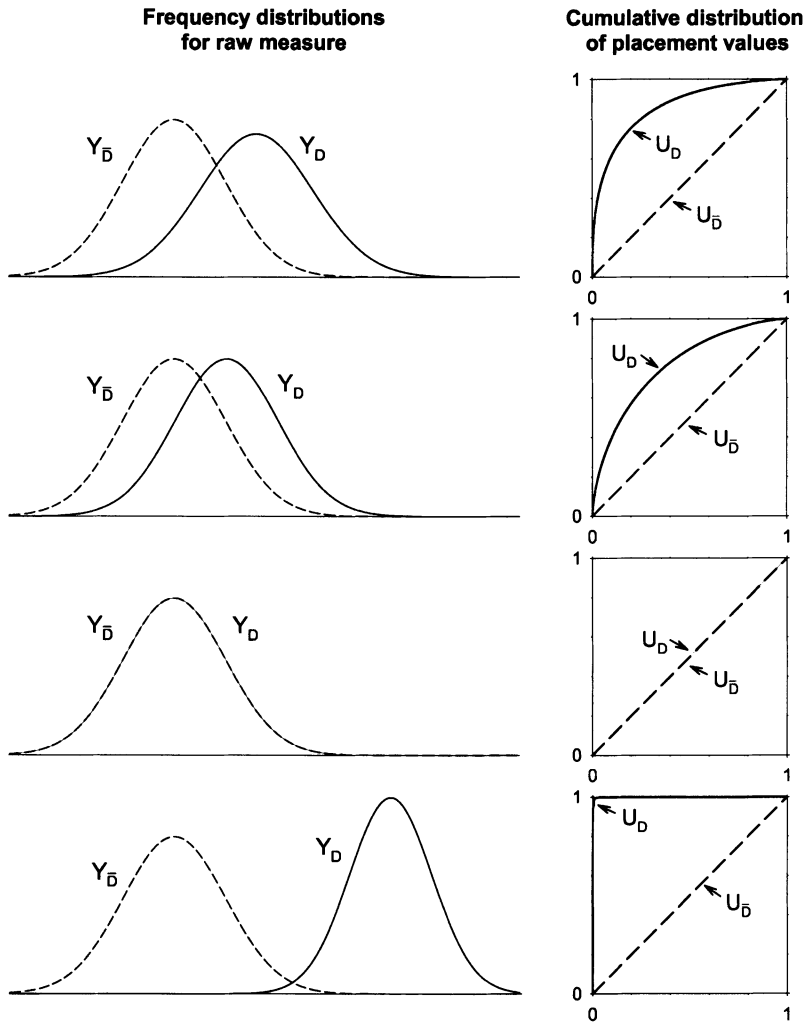

Figure 1. Distributions of the raw measure in the reference ($Y_{\bar{D}}$) and affected (Y_D) populations along with corresponding cumulative distributions for the placement values U_D and $U_{\bar{D}}$, respectively. Observe that monotone increasing transformations of Y yield the same placement values, so without loss of generality the distribution of $Y_{\bar{D}}$ is shown as N(0, 1).

covariate specific reference distribution. Let $\mathbf{Z}_D$ denote covariates that may affect discrimination in the sense that they are included in the model (2). Thus, for a subject from the affected population with covariates $\mathbf{X} = (\mathbf{Z}, \mathbf{Z}_D)$ and measure $Y_D, U_D \equiv 1 - F_{\bar{D},\mathbf{Z}}(Y_D)$, and we model $H_{\boldsymbol{\alpha}}(U_D) = -\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_D + \varepsilon$, where $\varepsilon \sim g(\cdot)$. Equivalently we can write

$$\mathrm{ROC}_{\mathbf{Z},\mathbf{Z}_D}(u) = g\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)\}.$$

The covariates $\mathbf{Z}$ and $\mathbf{Z}_D$ may have some or all components in common. The interpretation for $\mathrm{ROC}_{\mathbf{Z},\mathbf{Z}_D}$ is that it describes the separation between subjects in the reference population with covariates $\mathbf{Z}$ to those in the affected population with covariates $(\mathbf{Z}, \mathbf{Z}_D)$. Pepe (2003) describes several illustrative examples and presents methods for estimating the regression parameters with an algorithm proposed by Alonzo and Pepe (2002). We now propose a new algorithm that uses the likelihood of the placement values.

Since the distribution function for U_D is $P(U_D \le u \mid \mathbf{Z}_D) = g\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)\}$, its density function is

$$f_{\mathbf{Z}_D}(u) = \dot{g}\{\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)\}h_{\boldsymbol{\alpha}}(u),$$

where $\dot{g}(x) = (d/dx)g(x)$ and $h_{\boldsymbol{\alpha}}(u) = (d/du)\,H_{\boldsymbol{\alpha}}(u)$. Suppose for now that the reference distributions $F_{\bar{D},\mathbf{Z}}(\cdot)$ are known and that data for n_D random observations from the affected population are available. The log likelihood is $\Sigma \log f_{\mathbf{Z}_{Di}}(U_{Di})$. In many settings there is interest only in a portion of the ROC curve (Thompson and Zucchini, 1989; Dodd and Pepe, 2003b). In this case the regression model is specified only for false positive rates $u \in [a, b] \subset (0, 1)$, and the log likelihood is

$$\begin{aligned}\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n_D} \big[& I(U_{Di} < a) \log g\big(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{Di} + H_{\boldsymbol{\alpha}}(a)\big) \\ & + I(U_{Di} > b) \log\big\{1 - g\big(\boldsymbol{\beta}^{\mathrm{T}}\mathbf{Z}_{Di} + H_{\boldsymbol{\alpha}}(b)\big)\big\} \\ & + I(U_{Di} \in (a,b)) \log f_{\mathbf{Z}_{Di}}(U_{Di})\big], \end{aligned} \tag{3}$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}^{\mathrm{T}})$.

Thus far we have assumed that $F_{\bar{D},\mathbf{Z}}$ is known. In practice, data from $n_{\bar{D}}$ observations $\{(Y_{\bar{D}j}, \mathbf{Z}_j); j = 1, \ldots, n_{\bar{D}}\}$ will be available to estimate $F_{\bar{D},\mathbf{Z}}$. If $F_{\bar{D},\mathbf{Z}}$ is fully parameterized, then a joint likelihood can be used to estimate those parameters

and $\boldsymbol{\theta}$ simultaneously. The parameters estimated in this fashion are of course consistent and efficient. We choose instead to use semiparametric or preferably nonparametric methods for estimating $F_{\bar{D},\mathbf{Z}}$. The rationale is that ROC curves fundamentally describe the *relationship* between the distributions of the affected and reference populations, not the distributions themselves. They operate on a scale that is independent of the raw measurements of Y, and are, by definition, invariant to monotone transformations of Y. Procedures that fully parameterize the distribution of Y in the reference population impose more structure than is necessary given that the focus is only on the separation between distributions in the ROC metric. If $\mathbf{Z}$ is discrete it may be feasible to estimate the covariate specific reference distribution using the empirical cdfs for Y in the reference population. In other settings (e.g., when $\mathbf{Z}$ is continuous), a semiparametric approach can be taken, such as the location-scale model of Heagerty and Pepe (1999).

In our approach, the reference data provide estimates of the placement values

$$\hat{U}_{Di} = 1 - \hat{F}_{\bar{D},\mathbf{Z}i}(Y_{Di})$$

which are substituted into the log likelihood (3) to form a *pseudo-log-likelihood*, $l(\boldsymbol{\theta})$. It can be shown that $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\, l(\boldsymbol{\theta})$ is consistent and that $n_D^{1/2}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed (Pepe and Cai, 2002). Izmirlian (2003) takes another approach. He parameterizes the ROC curve using a proportional hazards model and estimates $F_{\bar{D},Z}(\cdot)$ with spline functions.

4. Numerical Studies

Simulation studies were conducted to investigate whether inference based on asymptotic theory is adequate for use with sample sizes likely to be encountered in practice. We generated independent observations for random samples from the affected and reference populations as

$$Y_D = \alpha_1^{-1}\{\alpha_0 + \beta_1 Z_1 + (\beta_2 + 0.5\alpha_1)Z_2 + \varepsilon_D\} \qquad (4)$$

and

$$Y_{\bar{D}} = 0.5Z_2 + \varepsilon_{\bar{D}}, \qquad (5)$$

where $\varepsilon_{\bar{D}}$ and ε_D are standard normal random variables and Z_1 and Z_2 are Bernoulli($p = 0.5$) and uniform(0, 1) random variables, respectively. The induced ROC curve is

$$\text{ROC}_{Z,Z_D}(u) = \Phi\big(\alpha_0 + \alpha_1\Phi^{-1}(u) + \beta_1 Z_1 + \beta_2 Z_2\big).$$

Observe that the covariate associated with Y in the reference population is $Z = Z_2$ while both Z_1 and Z_2 influence discrimination so that $\mathbf{Z}_D = (Z_1, Z_2)$.

The results shown in Table 1 indicate that inference based on the asymptotic theory appears to work rather well. As expected, estimation is more precise when the whole ROC curve is modeled (upper panel) compared with the setting where only the portion covering 20% of its domain is modeled (lower panel). Note also that in this latter case the estimated standard errors appear to be larger than the true standard errors with the smaller sample size ($n_D = n_{\bar{D}} = 100$). This yields confidence intervals with coverage that is somewhat higher than desired. Nevertheless, even in this case the results indicate that the method provides an adequate approach to inference.

Alonzo and Pepe (2002) propose an alternative algorithm for estimating the model parameters. They choose a finite set of values in $[a, b]$ denoted by $T = \{u_1, \ldots, u_{n_T}\}$ and calculate the n_T binary variables $B_{ui} = I[\hat{U}_{Di} \le u]$, $u \in T$ for $i = 1, \ldots, n_D$. Next, the binary generalized linear regression model

$$\text{E}\{B_{ui}\} = g\big\{\boldsymbol{\beta}^{\text{T}}\mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)\big\}$$

is fit to the data $\{(B_{ui}, \mathbf{Z}_D, h(u)), i = 1, \ldots, n_D, u = u_1, \ldots, u_{n_T}\}$, where $h(u)$ are the linear basis functions for $H_{\boldsymbol{\alpha}}$. In the simulation setting above, $H_{\boldsymbol{\alpha}}(u) = \alpha_0 + \alpha_1\,\Phi^{-1}(u)$, so that the basis functions are $(1, \Phi^{-1}(u))$. Standard marginal regression methods with independence working covariance matrix are used for fitting. The method gives valid consistent estimates because $E\,\{I[U_{Di} \le u]\} = P[U_{Di} \le u] = g\,\{\boldsymbol{\beta}^{\text{T}}\,\mathbf{Z}_D + H_{\boldsymbol{\alpha}}(u)\}$.

The performance of this approach to estimation was compared with that of the pseudo-likelihood approach using the same simulation models (4) and (5). The new pseudo-likelihood estimates are strikingly more efficient than the binary regression estimators, at least for estimating the regression coefficients β_1 and β_2 (Table 2). The mean-squared errors of estimators for β_2 using the pseudo-likelihood method

Table 1

Biases, empirical standard errors (ESE), mean estimated standard errors (MESE), and empirical 95% *coverage probabilities (ECP) of the pseudo-maximum-likelihood estimate of* ($\alpha_0 = 1.0$, $\alpha_1 = 1.0$, $\beta_1 = 0.5$, $\beta_2 = 0.7$). *The results are based on* 1000 *simulated data sets.*

	$n_D = 100$, $n_{\bar{D}} = 100$				$n_D = 200$, $n_{\bar{D}} = 400$			
	Bias	ESE	MESE	ECP	Bias	ESE	MESE	ECP
				(I) $a = 0.01$, $b = 0.99$				
α_0	0.004	0.274	0.271	0.952	−0.011	0.186	0.184	0.946
α_1	−0.032	0.166	0.164	0.940	−0.024	0.103	0.101	0.942
β_1	−0.019	0.427	0.433	0.951	0.010	0.276	0.275	0.951
β_2	0.006	0.222	0.223	0.960	0.009	0.153	0.150	0.947
				(II) $a = 0.01$, $b = 0.20$				
α_0	0.148	0.422	0.461	0.975	0.041	0.253	0.257	0.958
α_1	0.098	0.256	0.314	0.962	0.029	0.147	0.152	0.962
β_1	−0.016	0.424	0.479	0.967	0.007	0.291	0.286	0.948
β_2	0.001	0.220	0.223	0.957	0.013	0.160	0.155	0.939

Table 2

Estimates of $(\alpha_0, \alpha_1, \beta_1, \beta_2)$ *compared with their actual values, based on the pseudo-maximum-likelihood approach and on the binary regression approach. Results are based on* 1000 *simulated data sets and* $a = 0.01$.

			Pseudo-MLE				Binary regression			
			α_0	α_1	β_1	β_2	α_0	α_1	β_1	β_2
True value of the parameters: $\alpha_0 = 1.0$, $\alpha_1 = 1.0$, $\beta_1 = 0.5$, $\beta_2 = 0.7$										
$n_D = 50, n_{\bar{D}} = 50$	$b = 0.99$	Bias	0.045	0.015	0.031	0.020	0.096	0.130	0.031	0.042
		MSE	0.239	0.060	0.520	0.112	0.295	0.073	0.687	0.156
	$b = 0.50$	Bias	0.044	0.016	0.035	0.020	0.093	0.126	0.034	0.035
		MSE	0.260	0.073	0.523	0.114	0.306	0.089	0.641	0.147
$n_D = 50, n_{\bar{D}} = 100$	$b = 0.99$	Bias	0.139	0.097	−0.024	0.015	0.125	0.114	−0.005	0.036
		MSE	0.235	0.051	0.466	0.099	0.270	0.061	0.613	0.150
	$b = 0.50$	Bias	0.146	0.102	−0.022	0.015	0.114	0.101	−0.002	0.031
		MSE	0.250	0.060	0.471	0.100	0.275	0.072	0.597	0.137
True value of the parameters: $\alpha_0 = 1.5$, $\alpha_1 = 0.9$, $\beta_1 = 0.5$, $\beta_2 = 0.7$										
$n_D = 50, n_{\bar{D}} = 50$	$b = 0.99$	Bias	0.081	0.019	0.005	0.028	0.210	0.191	0.017	0.073
		MSE	0.314	0.085	0.535	0.140	0.581	0.132	0.862	0.250
	$b = 0.50$	Bias	0.073	0.016	0.009	0.028	0.189	0.172	0.017	0.064
		MSE	0.520	0.173	0.584	0.152	0.787	0.175	0.846	0.248
$n_D = 50, n_{\bar{D}} = 100$	$b = 0.99$	Bias	0.249	0.152	−0.009	0.012	0.236	0.173	0.004	0.074
		MSE	0.385	0.093	0.502	0.113	0.536	0.119	0.781	0.247
	$b = 0.50$	Bias	0.258	0.157	−0.007	0.012	0.213	0.151	0.001	0.063
		MSE	0.611	0.134	0.548	0.120	0.683	0.146	0.769	0.245

range from 46% to 77% of those from the binary regression approach, or a relative efficiency range of 1.3–2.2.

5. A Prognostic Marker in Cystic Fibrosis

The Cystic Fibrosis Foundation Registry gathers data annually from over 75% of cystic fibrosis patients in the United States. The disease is characterized by progressive deterioration of the lungs. The forced expiratory volume in one second (FEV_1) is considered a leading indicator of disease severity and is used as a prognostic indicator in these patients. Although their lungs are typically chronically infected, acute pulmonary exacerbations leading to hospitalization for intravenous antibiotic therapy are a major morbidity and the most common cause of death in these patients. Here we consider the degree to which FEV_1 can discriminate between patients who subsequently suffer a pulmonary exacerbation and those who do not.

For the analysis, we selected children between the ages of 6 and 18 years with data recorded in both 1995 and 1996 and for whom a routine throat culture was performed in 1995. Subjects were classified as having a pulmonary exacerbation in 1996 ($D = 1$) or not ($D = 0$). The discriminatory measure, Y, is FEV_1 measured in 1995 and calculated in the usual fashion as a percentage of that predicted for healthy non-CF children of the same age, height, and gender (Rosenfeld et al., 2001). Because higher values of Y are associated with better health and prognosis, we transformed it to $Y = -FEV_1$ to conform with our convention about higher values of Y being associated with $D = 1$. Covariates considered were age (in years), height (z-score; Hamill et al., 1997), and an indicator of whether or not the child's throat culture tested positive for *pseudomonas aeruginosa* (PA; coded as 1 for positive and 0 for negative), which is a chronic bacterial infection associated with poor prognosis for these patients. Although gender was also considered, it was not important for either the reference distribution model or the discrimination model.

The reference distribution model fit was

$$Y_{\bar{D}} = \gamma_1 \text{age} + \gamma_2 \text{height} + \gamma_3 \text{PA} + \varepsilon_{\bar{D}},$$

where $\varepsilon_{\bar{D}}$ has an unspecified distribution (Heagerty and Pepe, 1999). The ROC curve comparing FEV_1 in subjects who had pulmonary exacerbations the following year to those of the same age, height, and PA status who did not have exacerbations was modeled as

$$\text{ROC}(u) = \Phi\big(\alpha_0 + \alpha_1 \Phi^{-1}(u) + \beta_1 \text{age} + \beta_2 \text{height} + \beta_3 \text{PA}\big).$$

In terms of covariate specific placement values, the assumption is that

$$\Phi^{-1}(U_D) = -\beta_1 \text{age} - \beta_2 \text{height} - \beta_3 \text{PA} + \varepsilon_D,$$

where $\varepsilon_D \sim N(-\alpha_0, \alpha_1^{-1})$. Results are shown in Table 3. The discrimination achieved by FEV_1 in some specific age, gender, and height subpopulations is displayed in Figure 2.

Colonization with PA, increasing age, and poorer nutritional status (for which low height z-score is a surrogate measure) are all considered to be associated with poorer prognosis for CF patients. Not surprisingly, the analysis indicates that they are all associated with poorer lung function in the reference population and should be adjusted for in calculating placement values. Moreover, the parameters β_1, β_2, and β_3 indicate that the prognostic capacity of FEV_1 for discriminating between those who do and do not have a pulmonary exacerbation is superior in such patients. That is, FEV_1 is a better discriminator in older, less healthy patients and does not discriminate as well in younger, healthier patients.

Table 3

Fitted regression coefficients for FEV_1 as a predictor of pulmonary exacerbation in children with cystic fibrosis. Models were fit for $-FEV_1$. The pseudo-likelihood was used for estimating the discrimination model parameters.

	Reference population model			Discrimination model		
	Parameter	Estimate	SE	Parameter	Estimate	SE
Age (years-10)	γ_1	1.19	0.08	β_1	0.04	0.01
Height (z-score)	γ_2	−0.55	0.30	β_2	−0.44	0.05
PA (yes versus no)	γ_3	7.89	0.54	β_3	0.24	0.10
Location	–	–	–	α_0	0.96	0.09
Scale	–	–	–	α_1	1.59	0.05

6. Modeling the Mean Placement Value

Several authors have suggested that covariate effects on discrimination be evaluated with regression modeling of the AUC or other ROC summary index (Thompson and Zucchini, 1989; Dorfman, Berbaum, and Metz, 1992; Dodd and Pepe, 2003a).

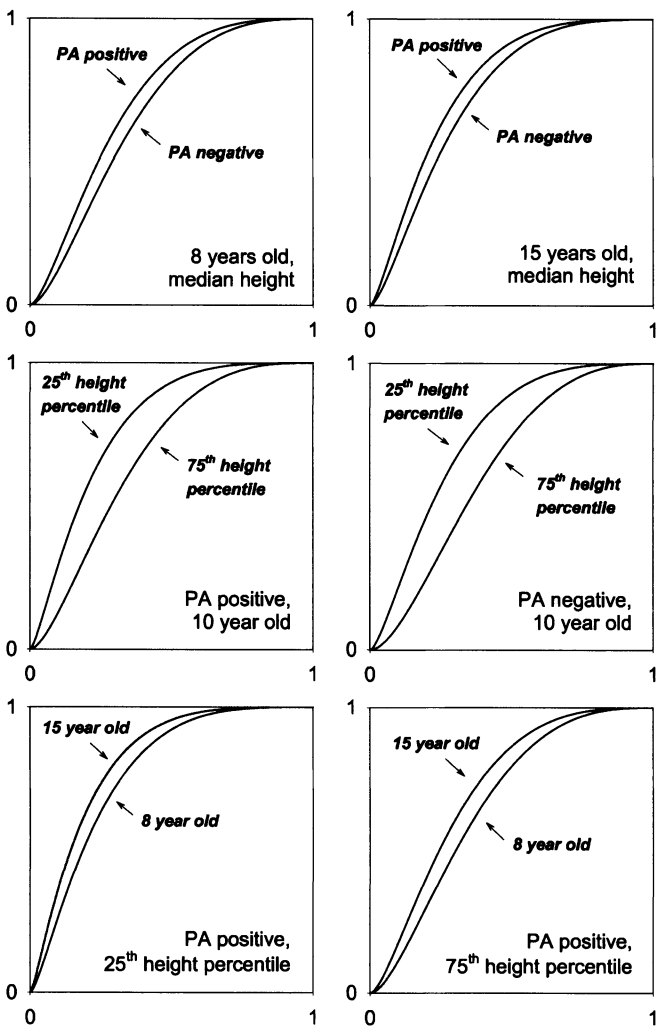

Figure 2. The fitted ROC curves, or equivalently the placement value cdfs, for FEV_1 in cystic fibrosis patients.

The most general formulation is

$$\mathrm{AUC}_{\mathbf{Z},\mathbf{Z}_D} = f\left(\eta_0 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}_D\right),$$

where $\mathrm{AUC}_{\mathbf{Z},\mathbf{Z}_D}$ is the area under the curve $\mathrm{ROC}_{\mathbf{Z},\mathbf{Z}_D}$. Earlier we noted the relationship between the AUC and the mean placement value, $E(U_D)$. Therefore, these regression models can be interpreted as models for the mean placement value. The above model is also written as

$$\mathrm{E}(1 - U_D \mid \mathbf{Z}_D) = f\left(\eta_0 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}_D\right),$$

and when viewed in this fashion, inferential procedures based on GLM iterative reweighted least squares fitting are suggested. The estimating equation is

$$\sum_{i=1}^{n_D} \begin{pmatrix} 1 \\ \mathbf{Z}_{Di} \end{pmatrix} w\left(\eta_0 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}_{Di}\right)\left\{1 - U_{Di} - f\left(\eta_0 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}_{Di}\right)\right\} = 0,$$

where $w(\eta_0 + \boldsymbol{\eta}^{\mathrm{T}}\mathbf{Z}_{Di})$ is a weight function. We substitute $\hat{U}_{Di}$ for U_{Di} in the usual setting where the reference distribution is estimated from data.

We fit the model

$$\mathrm{logit}(\mathrm{AUC}_{\mathbf{Z},\mathbf{Z}_D}) = \eta_0 + \eta_1 \mathrm{age} + \eta_2 \mathrm{height} + \eta_3 \mathrm{PA}$$

to the cystic fibrosis registry data, with weight function $w = 1$. Standard errors were calculated using the bootstrap. We find $\hat{\eta}_0 = 0.65\,(\mathrm{SE} = 0.073), \hat{\eta}_1 = 0.026\,(\mathrm{SE} = 0.011), \hat{\eta}_2 = -0.266\,(\mathrm{SE} = 0.039)$, and $\hat{\eta}_3 = 0.18\,(\mathrm{SE} = 0.084)$. We arrive at the same qualitative conclusions as arrived at earlier. FEV_1 is a better discriminator in older patients. Both colonization with PA and short stature are also associated with better ability to discriminate risk of pulmonary exacerbation using FEV_1. We refer to Dodd and Pepe (2003a) for more extensive discussion of regression modeling for AUCs and for interpretation of parameters in these models.

7. Discussion

This article proposes that in order to evaluate the discrimination achieved with a measure, one can look at the distribution of the standardized measure, namely the placement value, in the affected population. The more this distribution differs from uniform(0, 1) the better the discrimination achieved with the measure. The standardization provides a scale on which comparisons can be made across populations, across settings, and across measures that seek to discriminate between $D = 0$

and 1. Although this provides some motivation for assessing the placement value distribution, the fact that it is equal to the ROC curve that directly quantifies the operating characteristics of classification rules based on Y provides even more compelling motivation, in our opinion.

There is a lot of statistical methodology for making inference about distribution functions. In this article, we note that these methods can be adapted to yield new procedures for making inference about ROC curves by applying them to placement value distributions. The fact that the reference distribution is estimated in most applications implies that $\{\hat{U}_{D1}, \ldots, \hat{U}_{Dn_D}\}$ are not statistically independent. Asymptotic distribution theory is therefore more complicated than for classic applications in which observations are independent. In practice one can apply bootstrapping techniques, resampling data from both the reference and the affected samples to assess sampling variability of model parameters. We have derived distribution theory for the special case of pseudo-likelihood estimates of parameters in the ROC regression models of Pepe (1997).

Although the placement value distributions are continuous (because Y is), another complication that arises from estimating the reference distributions with nonparametric or semiparametric methods is that $\{\hat{U}_{D1}, \ldots, \hat{U}_{Dn_D}\}$ may have discrete mass points. For example, when the empirical cdf is used for $\hat{F}_{\bar{D}}$, placement values are zero for all $Y_{Di} > \max\{Y_{\bar{D}j}, j = 1, \ldots, n_{\bar{D}}\}$, a substantial proportion of observations if Y is a good discriminating measure. Transformations of the placement values that are required, such as $\Phi^{-1}(\hat{U}_D)$ in our examples, cannot be calculated for such values. We avoided this technical problem by restricting ROC modeling to a proper subinterval $[a, b]$ of (0, 1). Another justification is that in practice, it does not make sense to model at the ends of the (0, 1) range, because there is no overlap between the distributions of Y_D and $Y_{\bar{D}}$ in these regions. Nevertheless, further work to develop guidelines for dealing with the tails of the distribution of U_D would be welcome.

We mentioned that standardization to a reference distribution is a well-established concept in clinical practice and indeed beyond clinical medicine. Such standardization is typically done to induce comparability of measures across populations. We propose that the placement value standardization also provides comparability across different measures that are intended for the purpose of providing discrimination. This is already well known in some fields, namely those that already use ROC curves. We hope that the standardization concept will help to introduce ROC curves to other fields that seek to evaluate and compare discriminatory measures.

Acknowledgements

The authors are grateful for support provided by NIH grants GM-54438, CA-86368, and AI-29168.

Résumé

La médecine moderne utilise couramment des mesures dans un but de classement ou de prédiction (dosages de biomarqueurs, données cliniques, tests de biologie moléculaire, ...). L'évaluation scientifique de ces mesures implique d'apprécier leur capacité à prédire l'événement d'intérêt, le plus souvent à l'aide de courbes ROC. Ces courbes peuvent plus généralement se révéler utiles dans tout problème impliquant un classement en deux états ou deux populations (D = 0 ou D = 1). Nous montrons qu'une courbe ROC peut s'interpréter en termes de fonction de répartition de la mesure Y dans la population atteinte (D = 1) après standardisation par rapport à la distribution de Y dans la population de référence (D = 0). Nous qualifions les valeurs standardisées de ≪ valeurs de position ≫. Si les valeurs de position ont une distribution uniforme, Y n'est pas discriminante puisque sa distribution est la même dans les deux populations. Le degré auquel la distribution des valeurs standardisées diffère d'une loi uniforme sur [0, 1] fournit une manière naturelle de caractériser la capacité de discrimination de Y et suggère une interprétation non traditionnelle des courbes ROC. Les techniques utilisées pour effectuer des inférences sur les distributions suggèrent ainsi de nouvelles méthodes d'inférence sur les courbes ROC. Nous le montrons en considérant le modèle de régression GLM-ROC et en constatant qu'il est équivalant à un modèle de régression de la distribution des valeurs de position. La vraisemblance des valeurs de position fournit une nouvelle approche pour estimer les paramètres d'une courbe ROC, apparemment plus efficace que les méthodes antérieures. Nous appliquons cette méthode à l'évaluation d'une mesure de la fonction ventilatoire comme prédicteur de l'hospitalisation pour infection pulmonaire chez des patients atteints de mucoviscidose. Enfin, nous soulignons la relation entre la modélisation linéaire des valeurs de position et les modèles proposés récemment pour l'aire sous la courbe, l'indice le plus couramment utilisé pour mesurer la capacité de discrimination de Y.

References

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

Dodd, L. and Pepe, M. S. (2003a). Semi-parametric regression for the area under the receiver operating characteristic curve. *Journal of the American Statistical Association* **98,** 409–417.

Dodd, L. and Pepe, M. S. (2003b). Partial AUC estimation and regression. *Biometrics* **59,** 614–623.

Dorfman, D. D., Berbaum, K. S., and Metz, C. E. (1992). Receiver operating characteristic rating analysis: Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology* **27,** 723–731.

Elmer-Dewitt, P., Lemonick, M., Park, A., and Nash, M. (2001). Medicine: The future of drugs. *Time* **157,** 56–102.

Hamill, P. V., Drizd, T. A., Johnson, C. L., Reed, R. B., and Roche, A. F. (1997). NCHS growth curves for children birth–18 years. *Vital Health Statistics* **11,** 1–74.

Hanley, J. A. and Hajian-Tilaki, K. O. (1997). Sampling variability of non-parametric estimates of the areas under receiver operating characteristic curves: An update. *Academic Radiology* **4,** 49–58.

Heagerty, P. J. and Pepe, M. S. (1999). Semi-parametric estimation of regression quantiles with application to standardizing weight for height and age in US children. *Applied Statistics* **48,** 533–551; **31,** 227–237.

Izmirlian, G. (2003). A new efficient semiparametric family of models for the regression analysis of ROC curves. NCI Technical Report.

Pepe, M. S. (1997). A regression modeling framework for receiver operating characteristic curves in medical diagnostic testing. *Biometrika* **84,** 229–234.

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

Pepe, M. S. and Cai, T. (2002). *The analysis of placement values for evaluating discriminatory measures.* University of Washington Biostatistics Working Paper Series, paper 189.

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

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240,** 1285–1293.

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* New York: Academic Press.

Thompson, M. L. and Zucchini, W. (1989). On the statistical analysis of ROC curves. *Statistics in Medicine* **8,** 1277–1290.

van't Veer, L. J., Dai, H., van de Vijver, M. J., et al. (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415,** 530–536.

Zhou, X. H., McClish, D. K., and Obuchowski, N. A. (2002). *Statistical Methods in Diagnostic Medicine.* New York: Wiley.

Received March 2003. *Revised September* 2003.
Accepted November 2003.