UW Biostatistics Working Paper Series

9-30-2002

The Analysis of Placement Values for Evaluating Discriminatory Measures

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Tianxi Cai
Harvard University, tcai@hsph.harvard.edu

1 Introduction

This paper 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 (Veer *et al*, 2002) and predicting their response to therapy (The Chipping Forecast , 1999; Elmer-Dewitt *et al*, 2001).

It is important to quantify the discriminatory capacity of such measures before they are incorporated into the practice of medicine. Furthermore, measures must be compared formally in regards to their abilities to distinguish between populations so that the best discriminatory measures are identified. A statistical methodology for dealing with such questions is well developed in the context of evaluating diagnostic tests (Zhou, Obuchowski and McClish, 2002; Pepe, 2003). We and others have suggested that this so called ROC methodology should be adopted more broadly (Dodd and Pepe, 2002; Swets, 1986, 1988).

In the first part of this paper we provide a non-traditional view of ROC analysis. We hope that this new view point will make ROC analysis more accessible to mainstream statisticians and will clarify the potential for its broader applicability. The key idea is that the ROC curve is in fact the probability distribution of placement values, where placement values are defined as particular standardizations of the raw measurements relative to one of the populations. The distribution of the placement values quantifies in a natural way the separation between two populations and provides a mechanism to compare the discriminatory capacities of different measures. This leads us to propose that standard statistical methods for inference about probability distributions can be applied to the placement value random variables in order to make inference about ROC curves. We illustrate this in the second part of this paper.

First we consider ROC-GLM regression models (Pepe , 1997). In section 4 we show that consideration of the likelihood of placement values provides a new approach to making statistical inference about model parameters. Simulation studies in section 5 suggest that this approach is more efficient than existing methods of inference. Next, in section 7 we consider regression models for the area under the ROC curve (AUC). We show that they are equivalent to generalized linear models for the mean of the placement value random variable and suggest a related estimating equation. In sections 6 and 7 the approaches are applied to evaluate a measure of lung function for predicting acute pulmonary exacerbations in children with cystic fibrosis. In particular we consider factors that are associated with the discriminatory capacity of FEV_1 and find that it discriminates better in patients who are less healthy. The methods are also applied to a simple dataset involving the comparison of two biomarkers for cancer.

2 Placement values and the ROC curve.

Let Y be the continuous valued measure that is sought to discriminate between two populations. We use the indicator variable D to denote the populations, $D = 0$ or $D = 1$, and without loss of generality we assume larger values of Y are associated with $D = 1$ (otherwise Y can be transformed to achieve this). In the applications mentioned earlier, the populations could be (i) those diseased ($D = 1$) versus non-diseased ($D = 0$), in diagnostic testing; (ii) those who develop the serious condition within a certain time period or not, in screening; (iii) those who die of their illness versus those who do not, in prognostic research; and (iv) those who respond to a treatment versus non-responders, in therapeutic research. We choose one of the populations, namely $D = 0$, as the *reference* population. Let $F_{\bar{\mathrm{D}}}(y)$ denote the cumulative distribution function of Y in the reference population. We define the *placement value* to be

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

It is simply a transformation of Y that standardizes to the distribution in the reference population. The placement value is interpreted as the proportion of the reference population with values larger than Y. In essence it marks the placement of Y within the reference distribution. Interestingly, this sort of standardization is already commonly used 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 with the same age and gender (Hamill *et al*, 1977). If the child's weight is at the 90th percentile, then the equivalent placement value is 10%.

Let Y_{D} and $Y_{\bar{\mathrm{D}}}$ denote the measures from the populations with $D = 1$ and $D = 0$, respectively, and we call the former population the *affected* population to distinguish it from the reference population. The distribution of placement values in the reference population is uniform$(0, 1)$ by definition. On the other hand, the distribution of placement values in the affected population, i.e. the distribution of the random variable $U_{\mathrm{D}} = 1 - F_{\bar{\mathrm{D}}}(Y_{\mathrm{D}})$, 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 cumulative distribution function of U_{D} will not rise so steeply. In Figure 1 we illustrate both scenarios along with the two extreme cases where the distributions of Y_{D} and $Y_{\bar{\text{D}}}$ are identical and where they are completely separated.

[Figure 1 here]

The cumulative distribution of U_{D} is also known as the Receiver Operating Characteristic (ROC) curve. The ROC curve is commonly used to summarize the discriminatory capacity of a diagnostic test (Swets and Pickett , 1982; Pepe, 2000). It is usually motivated as a plot of the true versus false positive rates for classification rules based on Y, i.e. the plot of $1 - F_{\text{D}}(y) = P(Y > y \mid D = 1)$ versus $1 - F_{\bar{\text{D}}}(y) = P(Y > y \mid D = 0)$ for all threshold values $y \in (-\infty, \infty)$ that could be used to define test positivity. Setting $u = 1 - F_{\bar{\text{D}}}(y)$ so that $y = F_{\bar{\text{D}}}^{-1}(1-u)$ we see that this is the function

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

To see that the ROC curve is identical to the cumulative distribution function (cdf) for the placement values we write

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

Thus the ROC curve is really the distribution of Y in the affected population when Y is standardized relative to the reference population in a natural way (Figure 1).

The placement value standardization is particularly useful for comparing the discriminatory capacity of different measures. The standardization provides a common scale for measures that may be inherently incomparable on their original scales. For example, as predictors of long term outcome in AIDS patients, CD-4 counts and viral loads are incomparable as raw values (indeed they are measured in different units), but they can be compared when transformed to the placement value scale. In the upper two panels of Figure 2 we show the distributions of two biomarkers for pancreatic cancer in cancer cases and in controls (Wieand *et al*, 1989). It is hard to discern which provides the better discrimination. The distribution of the standardized values, i.e. the placement values, shown in the lower panel, indicates clearly and in a meaningful way that the CA 19-9 marker is superior in this respect. As another example, Pepe *et al* (2003) have recently argued that ROC curves are useful for comparing genes in regards to their differential expressions in two tissue types. Said another way, the distribution of gene expression values in the affected tissues after standardization to the distribution in control tissues provides a metric for comparing and ranking genes according to their differential expression between tissues.

[Figure 2 here]

When covariates $\mathbf{Z}$ affect the distribution of Y in the reference population, placement values are calculated with adjustment for such covariates. Let $F_{\bar{\mathrm{D}},\mathbf{Z}}$ denote the covariate specific reference distribution function. Then the placement value for a subject with covariates $\mathbf{Z}$ is

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

As mentioned earlier, in pediatric medicine, age and gender already are routinely adjusted for as covariates in standardizing anthropometric measures such as height and weight.

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. This happens for example when $\mathbf{Z}$ affects the location parameters for Y in the two populations in the same way as in the additive model:

$$Y = \alpha_0 + \boldsymbol{\alpha}_1^{\mathsf{T}}\mathbf{Z} + \alpha_2 D + \varepsilon, \quad \varepsilon \sim N(0,1).$$

In this case the placement values are defined to be $U = 1 - \Phi(Y - \alpha_0 - \boldsymbol{\alpha}_1^{\mathsf{T}}\mathbf{Z})$. For subjects in the affected population, $Y_{\mathrm{D}} - \alpha_0 - \boldsymbol{\alpha}_1^{\mathsf{T}}\mathbf{Z}$ has a normal distribution with mean α_2 and variance 1, and some algebra shows that the cdf of the transformed values, $U_{\mathrm{D}} = 1 - \Phi(Y_{\mathrm{D}} - \alpha_0 - \boldsymbol{\alpha}_1^{\mathsf{T}}\mathbf{Z})$ is $\Phi(\alpha_2 + \Phi^{-1}(u))$. That is $\mathrm{ROC}(u) = \Phi(\alpha_2 + \Phi^{-1}(u))$. In this particular example although the distributions of Y_{D} and $Y_{\bar{\mathrm{D}}}$ are affected by $\mathbf{Z}$, their separation is not because the distributions of U_{D} do not depend on covariates.

However, in practice, a measure might discriminate better in certain settings than in others. That is, covariates can affect the distribution of U_{D} in certain settings. For example, covariates that denote variations on the assay technique or protocol for ascertaining Y could be associated with variations in the discriminatory capacity of Y. Characteristics of study subjects can also affect the discriminatory capacity of Y. An example is age, which affects the capacity of mammography to distinguish between women with and without breast cancer. See Pepe (2003) and Dodd and Pepe (2002) for more examples.

Regression models for placement values in the affected population, U_{D}, can be used to quantify covariate effects on discrimination. One such model is

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

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

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

By writing the ROC regression model (2) as a model for the distribution of placement values (1), we have a new conceptual framework for the evaluation of covariates on discrimination. In addition, we will see in sections 4 and 7 that familiar tools such as maximum likelihood can be applied to make inference about covariate effects.

3 Current methodology with placement values.

It is interesting to consider some current commonly used statistical techniques in ROC analysis and to recast them in terms of procedures relating to placement values. Let $\{Y_{\bar{\mathrm{D}}1}, \ldots, Y_{\bar{\mathrm{D}}n_{\bar{\mathrm{D}}}}\}$ denote $n_{\bar{\mathrm{D}}}$ observations from the reference population and $\{Y_{\mathrm{D}1}, \ldots, Y_{\mathrm{D}n_{\mathrm{D}}}\}$ be a random sample from the affected population. The empirical cdf for $Y_{\bar{\mathrm{D}}}, \widehat{F}_{\bar{\mathrm{D}}}$, can be used to calculate empirical placement values

$$\widehat{U}_{\mathrm{D}i} = 1 - \widehat{F}_{\bar{\mathrm{D}}}(Y_{\mathrm{D}i}) = n_{\bar{\mathrm{D}}}^{-1} \sum_{j=1}^{n_{\bar{\mathrm{D}}}} I(Y_{\bar{\mathrm{D}}j} > Y_{\mathrm{D}i}),$$

where $\widehat{U}_{\mathrm{D}i}$ is interpreted as the proportion of reference observations that exceed $Y_{\mathrm{D}i}$. This transformation of Y_{D} is a non-parametric standardization to the reference distribution internal to the study. The empirical ROC curve is the standard non-parametric estimator of the ROC curve. Not surprisingly, it can be shown that the empirical distribution of $\{\widehat{U}_{\mathrm{D}1}, \ldots, \widehat{U}_{\mathrm{D}n_{\mathrm{D}}}\}$ is the empirical ROC curve. Thus Figure 2 shows the empirical ROC curves for two pancreatic cancer biomarkers.

The most common summary index of the ROC curve is the area under the curve (AUC). It is used as a summary descriptive statistic and as the basis for comparing ROC curves. Since 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 related to the AUC:

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

In addition, the sample average of $\{\widehat{U}_{\mathrm{D}1}, \ldots, \widehat{U}_{\mathrm{D}n_{\mathrm{D}}}\}$ is related to the standard non-parametric estimator of the AUC, which is well known to be the Mann-Whitney two-sample rank statistic:

$$\begin{aligned}
\widehat{\mathrm{AUC}} &\equiv \int \widehat{\mathrm{ROC}}(u)du = 1 - \int \left\{1 - \widehat{\mathrm{ROC}}(u)du\right\} du \\
&= 1 - n_{\mathrm{D}}^{-1} \sum_{i=1}^{n_{\mathrm{D}}} \widehat{U}_{\mathrm{D}i} = 1 - n_{\mathrm{D}}^{-1} n_{\bar{\mathrm{D}}}^{-1} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{j=1}^{n_{\bar{\mathrm{D}}}} I(Y_{\bar{\mathrm{D}}j} > Y_{\mathrm{D}i}) = n_{\mathrm{D}}^{-1} n_{\bar{\mathrm{D}}}^{-1} \sum_{i=1}^{n_{\mathrm{D}}} \sum_{j=1}^{n_{\bar{\mathrm{D}}}} I(Y_{\mathrm{D}i} \geq Y_{\bar{\mathrm{D}}j})
\end{aligned}$$

Traditionally two ROC curves are compared by calculating the difference between their estimated AUCs. Using an additional index for the two ROC curves we see that

$$\widehat{\mathrm{AUC}}_1 - \widehat{\mathrm{AUC}}_2 = n_{\mathrm{D}_2}^{-1} \sum_{i=1}^{n_{\mathrm{D}_2}} \widehat{U}_{\mathrm{D}_2 i} - n_{\mathrm{D}_1}^{-1} \sum_{i=1}^{n_{\mathrm{D}_1}} \widehat{U}_{\mathrm{D}_1 i}$$

The test statistic for comparing ROC curves is therefore based on the difference between sample means of the placement values. In other words, one first standardizes each of the two measures Y_1 and Y_2 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 current mainstream statistical methodology.

The interpretation of the non-parametric $\widehat{\mathrm{AUC}}$ as one minus the sample mean of placement values is not new. Indeed DeLong, DeLong and Clarke-Pearson (1988) use it in deriving asymptotic distribution theory. A more recent 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. Any statistic for testing equality of distribution functions therefore could be applied to the placement values in order to compare the ROC curves of two measures Y_1 and Y_2. The Kolmogorov-Smirnov statistic or the Wilcoxon rank sum statistic, for example, could be applied to compare ROC curves but have not yet been investigated for this purpose. We believe that the interpretation of the ROC curve as the distribution of placement values has not yet been fully exploited for the evaluation of discriminatory measures. There are many avenues to pursue. In the next sections of this paper we proceed to develop a procedure based on the likelihood of the placement values to make inference about covariate effects on discrimination.

4 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 $F_{\bar{\mathrm{D}},\mathbf{Z}}$ for the covariate specific reference distribution. Let $\mathbf{Z}_{\mathrm{D}}$ denote covariates that may affect discrimination in the sense that they are included in models for the placement value distribution of the affected population. Thus, for a subject from the affected population with covariates $\mathbf{X} = (\mathbf{Z}, \mathbf{Z}_{\mathrm{D}})$ and measure Y_{D}

$$U_{\mathrm{D}} \equiv 1 - F_{\bar{\mathrm{D}},\mathbf{Z}}(Y_{\mathrm{D}})$$

and we assume the model

$$H_{\boldsymbol{\alpha}}(U_{\mathrm{D}}) = -\boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{\mathrm{D}} + \varepsilon,$$

where $\varepsilon \sim g(\ )$. Equivalently we can write

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

The covariates $\mathbf{Z}$ and $\mathbf{Z}_\mathrm{D}$ may be the same or they may have only some components in common. The interpretation for $\mathrm{ROC}_{\mathbf{Z},\mathbf{Z}_\mathrm{D}}$ is that it describes the separation between subjects in the reference population with covariates $\mathbf{Z}$ to those in the affected population with covariates $(\mathbf{Z}, \mathbf{Z}_\mathrm{D})$.

Pepe (2003) describes several illustrative examples. To make the ideas concrete here consider the simple setting of Figure 2, where two biomarkers Y_1 and Y_2 are compared. The measurement Y represents Y_1 for some observations and Y_2 for others. The type of biomarker is a covariate, Z, with $Z = 0$ when Y represents Y_1 and $Z = 1$ when Y represents Y_2. This covariate clearly affects the distribution of Y. The covariate specific reference distribution function for Y is the distribution of Y_1 in the controls if $Z = 0$ and the distribution of Y_2 in controls if $Z = 1$. The biomarker type may also affect the discrimination capacity of Y. That is, the two markers may have different ROC curves. In our notation the reference and discrimination covariates are the same, $Z_\mathrm{D} = Z$, and the ROC curve $\mathrm{ROC}_Z(u)$ is the distribution of the placement values in the affected population for marker 1 when $Z = 0$ and that for marker 2 when $Z = 1$. A specific ROC model is

$$\mathrm{ROC}_Z(u) = \Phi\{\beta Z + \alpha_0 + \alpha_1\ \Phi^{-1}(u)\}.$$

Thus the ROC curve for marker 1 is represented as the binormal curve

$$\mathrm{ROC}_0(u) = \Phi\{\alpha_0 + \alpha_1\ \Phi^{-1}(u)\}$$

and that for marker 2 is

$$\mathrm{ROC}_1(u) = \Phi\{\alpha_0 + \beta + \Phi^{-1}(u)\}.$$

When the thresholds for positivity associated with the two markers are set to yield the same false positive rates, β quantifies the difference between the ROC curves on the probit scale. If β is positive, marker 2 is a better discriminator. For the same specificity it achieves better sensitivity. In terms of the distribution of placement values, this simple model implies that for marker 1, $\Phi^{-1}(U_\mathrm{D})$ has a normal distribution with (mean, sd) parameters $(-\alpha_0, \alpha_1^{-1})$ while the corresponding transform of the placement value for marker 2 has a normal distribution with parameters $(-\alpha_0 - \beta, \alpha_1^{-1})$. The ROC model quantifies the covariate effect as a location shift for the distribution of a transformation of the placement values.

Returning now to the general model formulation, we have that the distribution function for U_D is

$$P(U_\mathrm{D} \le u \mid \mathbf{Z}_\mathrm{D}) = g\{\boldsymbol{\beta}^\mathsf{T}\mathbf{Z}_\mathrm{D} + H_{\boldsymbol{\alpha}}(u)\}$$

which implies that the density function is

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

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 $F_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$ are known and that data for n_{D} random observations from the affected population are available. The log-likelihood is $\sum \log f_{\mathbf{Z}_{\mathrm{D}i}}(U_{\mathrm{D}i})$. In practice, one may only model a portion of the ROC curve, the regression model being specified only for $u \in [a, b] \ \subset (0, 1)$. Subject to this restricted model, the log-likelihood of the data is

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n_{\mathrm{D}}} \Big[I(U_{\mathrm{D}i} < a) \log g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_{\mathrm{D}i} + H_{\boldsymbol{\alpha}}(a)) + I(U_{\mathrm{D}i} > b) \log\{1 - g(\boldsymbol{\beta}^{\mathsf{T}}\mathbf{Z}_{\mathrm{D}i} + H_{\boldsymbol{\alpha}}(b))\}$$
$$+ I(U_{\mathrm{D}i} \in (a, b)) \log f_{\mathbf{Z}_{\mathrm{D}i}}(U_{\mathrm{D}i}) \Big] \qquad (3)$$

where $\boldsymbol{\theta} = (\boldsymbol{\alpha}, \boldsymbol{\beta}^{\mathsf{T}})$. Thus far we have assumed that $F_{\bar{\mathrm{D}},\mathbf{Z}}$ is known. In practice, data from $n_{\bar{\mathrm{D}}}$ observations $\{(Y_{\bar{\mathrm{D}}j}, \mathbf{Z}_j); j = 1 \ldots n_{\bar{\mathrm{D}}}\}$ will be available to estimate $F_{\bar{\mathrm{D}},\mathbf{Z}}$. If $F_{\bar{\mathrm{D}},\mathbf{Z}}$ 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 semi-parametric or preferably non-parametric methods for estimating $F_{\bar{\mathrm{D}},\mathbf{Z}}$. 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. Thus procedures that fully parameterize the distribution of Y in the reference population seem philosophically at odds with the concept of ROC curves. If $\mathbf{Z}$ is discrete (as in the two-marker example above) it may be feasible to estimate the covariate specific reference distribution using the empirical cdfs for Y in the reference population. Alternatively, and in particular when $\mathbf{Z}$ is continuous, the semi-parametric location-scale model of Heagerty and Pepe (1999) can be used. Therefore, in our approach, the reference data provide estimates of the placement values

$$\widehat{U}_{\mathrm{D}i} = 1 - \widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}_i}(Y_{\mathrm{D}i})$$

which are substituted into the log likelihood (3) to form a *pseudo log likelihood*, $\widehat{\ell}(\boldsymbol{\theta})$. We show in the appendix that $\widehat{\boldsymbol{\theta}} = \mathrm{argmax}_{\boldsymbol{\theta}}\, \widehat{\ell}(\boldsymbol{\theta})$ is consistent and that $n_{\mathrm{D}}^{1/2}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta})$ is asymptotically normally distributed.

5 Numerical studies.

Simulation studies were conducted to investigate if inference based on asymptotic theory is adequate for use with sample sizes likely to be encountered in practice. In addition, we compared the statistical efficiency of pseudo maximum likelihood estimates with the currently available methods for fitting ROC-GLM models.

5.1 Inference.

We generated independent observations for random samples from the affected and reference populations as

$$Y_{\mathrm{D}} = \alpha_1^{-1}\{\alpha_0 + \beta_1 Z_1 + (\beta_2 + 0.5\ \alpha_1) Z_2 + \varepsilon_{\mathrm{D}}\} \tag{4}$$

and

$$Y_{\bar{\mathrm{D}}} = 0.5\ Z_2 + \varepsilon_{\bar{\mathrm{D}}} \tag{5}$$

where $\varepsilon_{\bar{\mathrm{D}}}$ 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

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

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 $\mathbf{Z}_{\mathrm{D}} = (Z_1, Z_2)$.

The results shown in Table 1 indicate that inference based on the asymptotic theory of the appendix appears to work rather well for these data. 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 ($n_{\mathrm{D}} = n_{\bar{\mathrm{D}}} = 100$). This yields confidence intervals with coverage that is somewhat higher than desired. Nevertheless, even in this case the results indicate that the method provides a useful approach to inference.

5.2 Binary regression estimates.

Alonzo and Pepe (2002) propose the following algorithm for estimation that, interestingly, is also based on placement values: Choose a finite set of values in $[a, b]$ denoted by $T = \{u_1, \dots, u_{n_T}\}$ and calculate the n_T binary variables $B_{ui} = I[\widehat{U}_{\mathrm{D}i} \leq u] \quad u \in T$ for $i = 1, \dots, n_{\mathrm{D}}$. Next, fit the binary generalized linear regression model

$$E\{B_{ui}\} = g\{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{\mathrm{D}} + H_{\boldsymbol{\alpha}}(u)\}$$

to the data $\{(B_{ui}, \mathbf{Z}_{\mathrm{D}}, h(u)), i = 1, \dots n_{\mathrm{D}}, u = u_1, \dots u_{n_T}\}$, 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 $E\{I[U_{\mathrm{D}i} \leq u]\} = P[U_{\mathrm{D}i} \leq u] = g\{\boldsymbol{\beta}^{\mathsf{T}} \mathbf{Z}_{\mathrm{D}} + H_{\boldsymbol{\alpha}}(u)\}$.

The performance of this approach to estimation was compared with that of the pseudo-likelihood approach using the same simulation models (4) and (5). Results are shown in Table 2. The pseudo-likelihood estimates are strikingly more efficient than the binary regression estimators, at least for estimating the regression coefficients β_1 and β_2. The mean-squared errors of estimators for β_2 using the pseudo-likelihood method range from 46% to 77% of those from the binary regression approach, or a relative efficiency range of 1.3 to 2.2.

6 Applications.

6.1 Pancreatic cancer biomarkers.

We fit the following model to the pancreatic biomarker data displayed in Figure 2

$$\mathrm{ROC}_Z(u) = \Phi(\alpha_0 + \alpha_1 \Phi^{-1}(u) + \beta_1 Z + \beta_2 Z \Phi^{-1}(u)) \tag{6}$$

over the range $u \in (0, 0.2)$. We only consider the overlap between the distributions of Y_{D} and $Y_{\bar{\mathrm{D}}}$ in the upper 20% of the reference distributions. The model stipulates that in this region the placement values on a normal deviate scale, $\Phi^{-1}(U_{\mathrm{D}})$, have a $N(-\alpha_0, \alpha_1^{-1})$ distribution for marker 1 and a $N(-\alpha_0 - \beta_1, (\alpha_1 + \beta_2)^{-1})$ distribution for marker 2.

The model (6) was previously fit in Pepe (2000) using binary regression methods and yielded estimates for $\beta_1 = 0.23 (se = 0.71)$ and $\beta_2 = -0.91 (se = 0.46)$. The pseudo likelihood method yields similar results, $\beta_1 = 0.02 (se = 0.64)$ and $\beta_2 = -0.98 (se = 0.40)$. Note that for this data set the data are clustered in the sense that each subject contributes both a CA 19-9 marker measurement and a CA-125 marker measurement. We use the pseudo log-likelihood expression for estimation, although because the data are clustered we note that (3) is not a true log-likelihood and standard errors are calculated with the bootstrap. We are not surprised that the two methods are similar in their efficiencies in this example. Alonzo and Pepe (2002) show that the binary regression approach is reasonably efficient for fitting binormal ROC curves. The model (6) essentially only specifies that the ROC curves for the biomarkers are binormal. We therefore expect estimation that is already efficient with the binary regression method and so pseudo likelihood cannot improve upon it much.

6.2 A prognostic marker in cystic fibrosis patients.

The next example concerns registry data from the Cystic Fibrosis Foundation. The registry gathers data annually from over 75% of patients diagnosed with this genetic disorder in the United States. The disease is characterized by progressive deterioration of the lungs in these patients whose median survival is now quoted as approximately 30 years. 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 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 $Y = -\text{FEV}_1$ to conform with our convention about higher values of Y being associated with $D = 1$. Covariates considered were age, gender, height and an indicator of whether or not the child's throat culture tested positive for pseudomonas aeruginosa (PA) which is a chronic bacterial infection associated with poor prognosis for these patients. It turned out that gender was not an important covariate for either the reference distribution model or the discrimination model so we report results only for the covariates: age (in years), height (z-score, Hamill *et al* , 1997), and PA (coded as 1 for positive and 0 for negative).

The reference distribution model fit was

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

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

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

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

$$\Phi^{-1}(U_{\mathrm{D}}) = -\beta_1 \text{ age } - \beta_2 \text{ height } - \beta_3 \text{ PA } + \varepsilon_{\mathrm{D}}$$

where $\varepsilon_{\mathrm{D}} \sim N(-\alpha_0, \alpha_1^{-1})$. Results are shown in Table 3. The discrimination achieved by FEV_1 in some age, gender and height specific subpopulations is displayed in the ROC curves (or placement value cdfs) of Figure 3.

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. The prognostic capacity of FEV_1 for discriminating between those who do and do not have a pulmonary exacerbation appears to be 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.

7 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; Pepe, 1998; and Dodd and Pepe, 2003). The most general formulation is

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

where $\mathrm{AUC}_{\mathbf{Z},\mathbf{Z}_{\mathrm{D}}}$ is the area under the curve $\mathrm{ROC}_{\mathbf{Z},\mathbf{Z}_{\mathrm{D}}}$. Earlier we noted the relationship between the AUC and the mean placement value, $E(U_{\mathrm{D}})$. Therefore, these regression models can be interpreted as models for the mean placement value. In particular,

$$E(1 - U_{\mathrm{D}} \mid \mathbf{Z}_{\mathrm{D}}) = f(\eta_0 + \boldsymbol{\eta}^{\mathsf{T}} \mathbf{Z}_{\mathrm{D}}),$$

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

$$\sum_{i=1}^{n_{\mathrm{D}}} \begin{pmatrix} 1 \\ \mathbf{Z}_{\mathrm{D}i} \end{pmatrix} w(\eta_0 + \boldsymbol{\eta}^{\mathsf{T}} \ \mathbf{Z}_{\mathrm{D}i})\{1 - U_{\mathrm{D}i} - f(\eta_0 + \boldsymbol{\eta}^{\mathsf{T}} \mathbf{Z}_{\mathrm{D}i})\} = 0$$

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

To illustrate, let's return to the cystic fibrosis registry data. We fit the model

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

using the estimating equation above, with weight function $w = 1$, where the reference distribution for calculating $\widehat{U}_{\mathrm{D}i}$ was based on the same semi-parametric regression model as in section 6.2. Standard errors were calculated using the bootstrap. We find $\hat{\eta}_0 = 0.65(se = 0.073), \hat{\eta}_1 = 0.026(se = 0.011), \hat{\eta}_2 = -0.266(se = 0.039)$, and $\hat{\eta}_3 = 0.18(se = 0.084)$. 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 (2002) for discussion of the interpretation of the parameters $(\gamma_1, \gamma_2, \gamma_3)$ in these models in terms of log odds ratios for "correct ordering probabilities."

8 Discussion.

This paper proposes that to evaluate the discrimination achieved with a measure, one should look at the distribution of the placement value standardized measure 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 $D = 1$. 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 and these methods can be adapted for inference about placement value distributions. The fact that the reference distribution is estimated in most applications implies that $\{\widehat{U}_{\mathrm{D}1}, \ldots, \widehat{U}_{\mathrm{D}n_{\mathrm{D}}}\}$ 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, re-sampling data from both the reference and the affected samples to assess sampling variability of model parameters.

Although the placement value distributions are continuous (because Y is), another complication that arises from estimating the reference distributions with non-parametric or semi-parametric methods is that $\{\widehat{U}_{\mathrm{D}1}, \ldots, \widehat{U}_{\mathrm{D}n_{\mathrm{D}}}\}$ may have discrete mass points. For example, when the empirical cdf is used for $\hat{F}_{\bar{D}}$, placement values are zero for all $Y_{\text{D}i} > \max\{Y_{\text{D}j}, j = 1, \ldots, n_{\text{D}}\}$, a substantial proportion of observations if Y is a good discriminating measure. Transformations of the placement values that are required, such as $\Phi^{-1}(\widehat{U}_{\text{D}})$ 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 doesn't 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_{D} in these regions.

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 discriminatory measures.

Acknowledgments

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

References

[1] Alonzo, T.A. and Pepe, M.S. (2000). Distribution-free analysis using binary regression techniques. *Biostatistics* ,(in press).

[2] DeLong, E.R., DeLong, D.M., and Clarke-Pearson, D.L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a non-parametric approach. *Biometrics* **44**, 837-845.

[3] Dodd L. and Pepe, M.S. (2002). Semi-parametric regression for the area under the Receiver Operating Characteristic Curve. *Journal of the American Statistical Association*, (accepted subject to revision).

[4] 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.

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

[6] 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.

[7] 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.

[8] 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.

[9] 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.

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

[11] Pepe, M.S. (2000). Statistics in the life and medical sciences: a vignette on ROC methodology. *Journal of the American Statistical Association* **95**, 308-311.

[12] Pepe, M.S. (2003). *Statistical Evaluation of Diagnostic Tests and Biomarkers.* Oxford University Press.

[13] Pepe, M.S., Longton, G., Anderson, G., Schummer, M. (2003). Selecting differentially expressed genes from micro-array experiments. *Biometrics*, (in press).

[14] Serfling, R.J. (1980). Approximation Theorems of Mathematical Statistics. Wiley.

[15] Swets, J.A. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin* **99**, 100-117.

[16] Swets, J.A. (1988). Measuring the accuracy of diagnostic systems. *Science* **240**, 1285-1293.

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

[18] The Chipping Forecast (1999). *Nature Genetics* **21**, supplement.

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

[20] van 'T Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Hart, A.A., *et al.* (2002). Gene expression profiling predicts clinical outcome of breast cancer. *Nature* **415**, 530-536.

[21] Wieand, S., Gail, M.H., James, B.R., and James, K.L. (1989). A family of non-parametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585-592.

[22] Zhou, X.H., McClish, D.K., and Obuchowski, N.A. (2002). *Statistical Methods in Diagnostic Medicine.* Wiley.

Appendix

To show the consistency and asymptotic normality of $\widehat{\boldsymbol{\theta}}$, we assume that the estimators of $F_{\bar{\mathrm{D}},\mathbf{Z}}(y)$ are uniformly consistent and $n_{\bar{\mathrm{D}}}^{\frac{1}{2}}\left\{\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(y) - F_{\bar{\mathrm{D}},\mathbf{z}}(y)\right\}$ converges to a Gaussian process uniformly in y and $\mathbf{Z}$. Without loss of generality, we also assume that $n_{\bar{\mathrm{D}}}^{\frac{1}{2}}\left\{\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(y) - F_{\bar{\mathrm{D}},\mathbf{Z}}(y)\right\}$ can be approximated by a sum of independent terms:

$$\sup_{y,\mathbf{Z}}\left| n_{\bar{\mathrm{D}}}^{\frac{1}{2}}\left\{\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(y) - F_{\bar{\mathrm{D}},\mathbf{Z}}(y)\right\} - n_{\bar{\mathrm{D}}}^{-\frac{1}{2}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}} I_{\bar{\mathrm{D}}j}(y,\mathbf{Z})\right| \to 0 \tag{7}$$

in probability. The estimator based on the location-scale model of Heagerty and Pepe (1999) satisfies these conditions (see Cai & Pepe, 2002 for derivations of the asymptotics for this estimator). If there is no covariate, the empirical estimate of $F_{\bar{\mathrm{D}}}(y)$ satisfies this condition as well.

Suppose that $\boldsymbol{\theta}_0 = (\boldsymbol{\alpha}_0, \boldsymbol{\beta}_0)$, the true value of $\boldsymbol{\theta}$, lies in a compact set $\boldsymbol{\mathcal{D}}_{\boldsymbol{\theta}}$ and the covariates are bounded. To show the consistency of $\widehat{\boldsymbol{\theta}}$, it is sufficient to show that $n_{\mathrm{D}}^{-1}\widehat{\ell}(\boldsymbol{\theta})$ converges uniformly to a deterministic function which has a unique maximizer at $\boldsymbol{\theta}_0$ (Newey and McFadden, 1994). It follows from the standard likelihood theory that $n_{\mathrm{D}}^{-1}\ell(\boldsymbol{\theta})$ has a unique maximizer and $\mathrm{argmax}_{\boldsymbol{\theta}}\, \ell(\boldsymbol{\theta})$ is a consistent estimate of $\boldsymbol{\theta}_0$. It follows from the uniform consistency of $\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}$ that

$$n_{\mathrm{D}}^{-1}\left\{\widehat{\ell}(\boldsymbol{\theta}) - \ell(\boldsymbol{\theta})\right\} \to 0$$

uniformly in $\boldsymbol{\theta} \in \boldsymbol{\mathcal{D}}_{\theta}$. This concludes the consistency of $\widehat{\boldsymbol{\theta}}$.

Let $\widehat{\mathbf{V}}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\widehat{\ell}(\boldsymbol{\theta})$ and $\mathbf{V}(\boldsymbol{\theta}) = \frac{\partial}{\partial\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$. To show that $n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0)$ is asymptotically normal, we take a Taylor series expansion of $\widehat{\mathbf{V}}(\widehat{\boldsymbol{\theta}})$ around $\boldsymbol{\theta}_0$ and obtain

$$n_{\mathrm{D}}^{\frac{1}{2}}(\widehat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \approx \left\{-n_{\mathrm{D}}^{-1}\frac{\partial\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)}{\partial\boldsymbol{\theta}}\right\}^{-1}\left\{n_{\mathrm{D}}^{-\frac{1}{2}}\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)\right\}.$$

It follows from the strong law of large numbers and the uniform consistency of $\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(\cdot)$ that $-n_{\mathrm{D}}^{-1}\frac{\partial}{\partial\boldsymbol{\theta}}\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)$ converges to $\mathbb{A}$. It remains to show that $n_{\mathrm{D}}^{-\frac{1}{2}}\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)$ converges to a normal distribution.

To this end, let $\dot{g}(x) = \frac{d}{dx}g(x)$, $\xi_f(\cdot) = \frac{d}{dx}\log f(x)$ for any function $f(\cdot)$, $\dot{\mathbf{H}}_{\boldsymbol{\alpha}}(u) = \frac{\partial}{\partial\boldsymbol{\alpha}}H_{\boldsymbol{\alpha}}(u)$, $\dot{\mathbf{h}}_{\boldsymbol{\alpha}}(u) = \frac{\partial}{\partial\boldsymbol{\alpha}}h_{\boldsymbol{\alpha}}(u)$, $e_i(u) = H_{\boldsymbol{\alpha}_0}(u) + \boldsymbol{\beta}_0^{\mathsf{T}}\mathbf{Z}_{\mathrm{D}i}$, $\delta_{0i} = I(U_{\mathrm{D}i} > b)$, $\delta_{1i} = I(U_{\mathrm{D}i} \in [a,b]$, $\delta_{2i} = I(U_{\mathrm{D}i} < a)$,

$$\mathbf{E}_{2i} = \xi_g\{e_i(a)\}\begin{bmatrix}\dot{\mathbf{H}}_{\boldsymbol{\alpha}_0}(a)\\ \mathbf{Z}_{\mathrm{D}i}\end{bmatrix}, \qquad \mathbf{E}_{0i} = \xi_{1-g}\{e_i(b)\}\begin{bmatrix}\dot{\mathbf{H}}_{\boldsymbol{\alpha}_0}(b)\\ \mathbf{Z}_{\mathrm{D}i}\end{bmatrix}, \qquad \mathbf{E}_i(u) = \xi_{\dot{g}}\{e_i(u)\}\begin{bmatrix}\dot{\mathbf{H}}_{\boldsymbol{\alpha}_0}(u)\\ \mathbf{Z}_{\mathrm{D}i}\end{bmatrix} + \begin{bmatrix}\frac{\dot{\mathbf{h}}_{\boldsymbol{\alpha}_0}(u)}{h_{\boldsymbol{\alpha}_0}(u)}\\ \mathbf{0}\end{bmatrix}.$$

To derive the large sample distribution of $n_{\mathrm{D}}^{\frac{1}{2}}\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)$, we write

$$\begin{aligned} n_{\mathrm{D}}^{\frac{1}{2}}\left\{\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)-\mathbf{V}(\boldsymbol{\theta}_0)\right\} =& n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\delta_{1i}\left\{\mathbf{E}_i(\widehat{U}_{\mathrm{D}i})-\mathbf{E}_i(U_{\mathrm{D}i})\right\} \\ &+n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\left\{(\widehat{\delta}_{1i}-\delta_{1i})\mathbf{E}_i(\widehat{U}_{\mathrm{D}i})+(\widehat{\delta}_{2i}-\delta_{2i})\mathbf{E}_{2i}+(\widehat{\delta}_{0i}-\delta_{0i})\mathbf{E}_{0i}\right\}. \end{aligned}$$

where $\widehat{\delta}_{ki}$ is δ_{ki} with $U_{\mathrm{D}i}$ replaced by $\widehat{U}_{\mathrm{D}i}$, for $k=0,1,2$. Since g has continuous third derivative and the covariates are bounded, we have $\dot{\mathbf{E}}_i(u)=\frac{d}{du}\mathbf{E}_i(u)$ bounded in $[a,b]$. This, coupled with the weak convergence of $n_{\bar{\mathrm{D}}}^{\frac{1}{2}}\left\{\widehat{F}_{\bar{\mathrm{D}},\mathbf{Z}}(y)-F_{\bar{\mathrm{D}},\mathbf{Z}}(y)\right\}$, ensures that

$$n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\delta_{1i}\left\{\mathbf{E}_i(\widehat{U}_{\mathrm{D}i})-\mathbf{E}_i(U_{\mathrm{D}i})\right\}\approx n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\delta_{1i}\dot{\mathbf{E}}_i(U_{\mathrm{D}i})(\widehat{U}_{\mathrm{D}i}-U_{\mathrm{D}i})\approx n_{\mathrm{D}}^{-\frac{1}{2}}n_{\bar{\mathrm{D}}}^{-1}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\mathbf{V}_{\bar{\mathrm{D}}1ij}$$

where $\mathbf{V}_{\bar{\mathrm{D}}1ij}=\delta_{1i}\dot{\mathbf{E}}_i(U_{\mathrm{D}i})I_{\bar{\mathrm{D}}j}(Y_{\mathrm{D}i},\mathbf{Z}_i)$.

It follows from the equicontinuity of $n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\left[I\left\{Y_{\mathrm{D}i}\geq c\right\}-\mathrm{ROC}_{\mathbf{Z}_i,\mathbf{Z}_{\mathrm{D}i}}\left\{1-F_{\bar{\mathrm{D}},\mathbf{Z}_i}(c)\right\}\right]\mathbf{E}_{ki}$ that

$$n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\left(\widehat{\delta}_{2i}-\delta_{2i}\right)\mathbf{E}_{2i}\approx n_{\mathrm{D}}^{-\frac{1}{2}}n_{\bar{\mathrm{D}}}^{-1}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\mathbf{E}_{2i}I_{\bar{\mathrm{D}}j}\left\{F_{\bar{\mathrm{D}},\mathbf{Z}_i}^{-1}(1-a),\mathbf{Z}_i\right\}f_{\mathbf{Z}_{\mathrm{D}i}}(a),$$

and

$$n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\left(\widehat{\delta}_{0i}-\delta_{0i}\right)\mathbf{E}_{0i}\approx n_{\mathrm{D}}^{-\frac{1}{2}}n_{\bar{\mathrm{D}}}^{-1}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}}\mathbf{E}_{0i}I_{\bar{\mathrm{D}}j}\left\{F_{\bar{\mathrm{D}},\mathbf{Z}_i}^{-1}(1-b),\mathbf{Z}_i\right\}f_{\mathbf{Z}_{\mathrm{D}i}}(b).$$

Similarly, the equicontinuity of $n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\left[I(Y_{\mathrm{D}i}\in[c_1,c_2])\mathbf{E}_i(U_{\mathrm{D}i})-E\left\{I(Y_{\mathrm{D}i}\in[c_1,c_2])\mathbf{E}_i(U_{\mathrm{D}i})\mid\mathbf{Z}_i,\mathbf{Z}_{\mathrm{D}i}\right\}\right]$ ensures that

$$\begin{aligned} n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\left(\widehat{\delta}_{1i}-\delta_{1i}\right)\mathbf{E}_i(U_{\mathrm{D}i})\approx n_{\mathrm{D}}^{-\frac{1}{2}}n_{\bar{\mathrm{D}}}^{-1}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}}&\left\{\mathbf{E}_i(b)I_{\bar{\mathrm{D}}j}\left\{F_{\bar{\mathrm{D}},\mathbf{Z}_i}^{-1}(1-b),\mathbf{Z}_i\right\}f_{\mathbf{Z}_{\mathrm{D}i}}(b)\right. \\ &\left.-\mathbf{E}_i(a)I_{\bar{\mathrm{D}}j}\left\{F_{\bar{\mathrm{D}},\mathbf{Z}_i}^{-1}(1-a),\mathbf{Z}_i\right\}f_{\mathbf{Z}_{\mathrm{D}i}}(a)\right\} \end{aligned}$$

Let $\mathbf{V}_{\mathrm{D}i}=\delta_{1i}\mathbf{E}_{1i}+\delta_{2i}\mathbf{E}_{2i}+\delta_{0i}\mathbf{E}_{0i}$ and

$$\begin{aligned} \mathbf{V}_{\bar{\mathrm{D}}2ij}=&\left\{\mathbf{E}_{2i}-\mathbf{E}_i(a)\right\}I_{\bar{\mathrm{D}}j}\left\{F_{\bar{\mathrm{D}},\mathbf{Z}_i}^{-1}(1-a),\mathbf{Z}_i\right\}f_{\mathbf{Z}_{\mathrm{D}i}}(a) \\ &-\left\{\mathbf{E}_{0i}-\mathbf{E}_i(b)\right\}I_{\bar{\mathrm{D}}j}\left\{F_{\bar{\mathrm{D}},\mathbf{Z}_i}^{-1}(1-b),\mathbf{Z}_i\right\}f_{\mathbf{Z}_{\mathrm{D}i}}(b). \end{aligned}$$

Then $n_{\mathrm{D}}^{\frac{1}{2}}\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)$ is asymptotically equivalent to $n_{\mathrm{D}}^{-\frac{1}{2}}\sum_{i=1}^{n_{\mathrm{D}}}\mathbf{V}_{\mathrm{D}i}+n_{\mathrm{D}}^{-\frac{1}{2}}n_{\bar{\mathrm{D}}}^{-1}\sum_{i=1}^{n_{\mathrm{D}}}\sum_{j=1}^{n_{\bar{\mathrm{D}}}}(\mathbf{V}_{\bar{\mathrm{D}}1ij}+\mathbf{V}_{\bar{\mathrm{D}}2ij})$ which is a U-statistics (Serfling, 1980). It follows from the asymptotic properties of U-statistics that $n_{\mathrm{D}}^{\frac{1}{2}}\widehat{\mathbf{V}}(\boldsymbol{\theta}_0)$ converges in distribution to a normal random vector.

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 datasets.

(I) $a = 0.01$, $b = 0.99$.

	$n_{\mathrm{D}} = 100$, $n_{\bar{\mathrm{D}}} = 100$				$n_{\mathrm{D}} = 200$, $n_{\bar{\mathrm{D}}} = 400$			
	Bias	ESE	MESE	ECP	Bias	ESE	MESE	ECP
α_0	.004	.274	.271	.952	-.011	.186	.184	.946
α_1	-.032	.166	.164	.940	-.024	.103	.101	.942
β_1	-.019	.427	.433	.951	.010	.276	.275	.951
β_2	.006	.222	.223	.960	.009	.153	.150	.947

(II) $a = 0.01$, $b = 0.20$,

	$n_{\mathrm{D}} = 100$, $n_{\bar{\mathrm{D}}} = 100$				$n_{\mathrm{D}} = 200$, $n_{\bar{\mathrm{D}}} = 400$			
	Bias	ESE	MESE	ECP	Bias	ESE	MESE	ECP
α_0	.148	.422	.461	.975	.041	.253	.257	.958
α_1	.098	.256	.314	.962	.029	.147	.152	.962
β_1	-.016	.424	.479	.967	.007	.291	.286	.948
β_2	.001	.220	.223	.957	.013	.160	.155	.939

Table 2: Estimates of $(\alpha_0, \alpha_1, \beta_1, \beta_2)$ compared with their actual values, based on the pseudo maximum likelihood (Pseudo-MLE) approach and on the binary regression approach. Results are based on 1000 simulated datasets and $a = 0.01$.

(I) True value of the parameters: $\alpha_0 = 1.0, \alpha_1 = 1.0, \beta_1 = 0.5, \beta_2 = 0.7$.

			Psuedo-MLE				GEE			
			α_0	α_1	β_1	β_2	α_0	α_1	β_1	β_2
$n_{\text{D}} = 50,\ n_{\bar{\text{D}}} = 50$	$b = .99$	Bias	.045	.015	.031	.020	.096	.130	.031	.042
		MSE	.239	.060	.520	.112	.295	.073	.687	.156
	$b = .50$	Bias	.044	.016	.035	.020	.093	.126	.034	.035
		MSE	.260	.073	.523	.114	.306	.089	.641	.147
$n_{\text{D}} = 50,\ n_{\bar{\text{D}}} = 100$	$b = .99$	Bias	.139	.097	$-.024$	.015	.125	.114	$-.005$	.036
		MSE	.235	.051	.466	.099	.270	.061	.613	.150
	$b = .50$	Bias	.146	.102	$-.022$	.015	.114	.101	$-.002$	.031
		MSE	.250	.060	.471	.100	.275	.072	.597	.137

(II) True value of the parameters: $\alpha_0 = 1.5, \alpha_1 = 0.9, \beta_1 = 0.5, \beta_2 = 0.7$.

			Psuedo-MLE				GEE			
			α_0	α_1	β_1	β_2	α_0	α_1	β_1	β_2
$n_{\text{D}} = 50,\ n_{\bar{\text{D}}} = 50$	$b = .99$	Bias	.081	.019	.005	.028	.210	.191	.017	.073
		MSE	.314	.085	.535	.140	.581	.132	.862	.250
	$b = .50$	Bias	.073	.016	.009	.028	.189	.172	.017	.064
		MSE	.520	.173	.584	.152	.787	.175	.846	.248
$n_{\text{D}} = 50,\ n_{\bar{\text{D}}} = 100$	$b = .99$	Bias	.249	.152	$-.009$	.012	.236	.173	.004	.074
		MSE	.385	.093	.502	.113	.536	.119	.781	.247
	$b = .50$	Bias	.258	.157	$-.007$	.012	.213	.151	.001	.063
		MSE	.611	.134	.548	.120	.683	.146	.769	.245

Table 3: Fitted regression co-efficients for FEV_1 as a predictor of pulmonary exacerbation in children with cystic fibrosis. Models were fit for $-\mathrm{FEV}_1$. 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	0.009
height (z-score)	γ_2	-0.55	0.30	β_2	-0.44	0.051
PA (yes versus no)	γ_3	7.89	0.54	β_3	0.24	0.101
location	-	-	-	α_0	0.96	0.091
scale	-	-	-	α_1	1.59	0.050

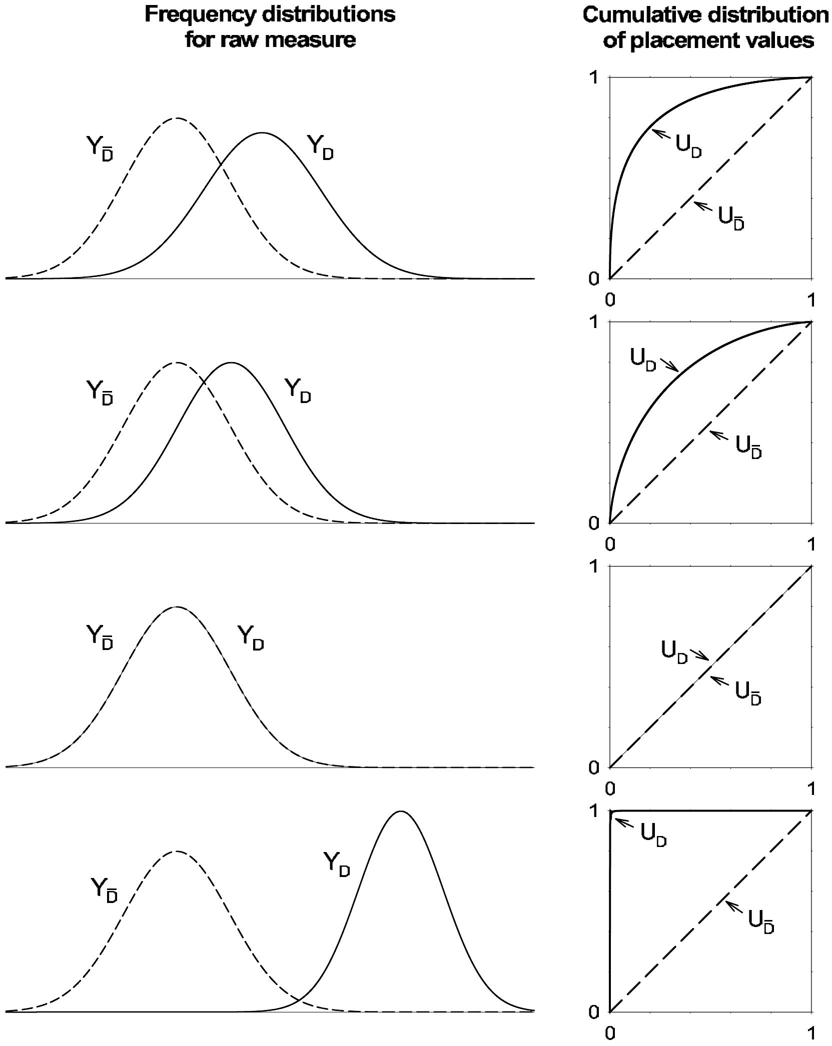

Figure 1: Distributions of the raw measure in the reference $(Y_{\bar{\mathrm{D}}})$ and affected (Y_{D}) populations along with corresponding cumulative distributions for the placement values U_{D} and $U_{\bar{\mathrm{D}}}$, respectively. Observe that the monotone increasing transformations of Y yield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

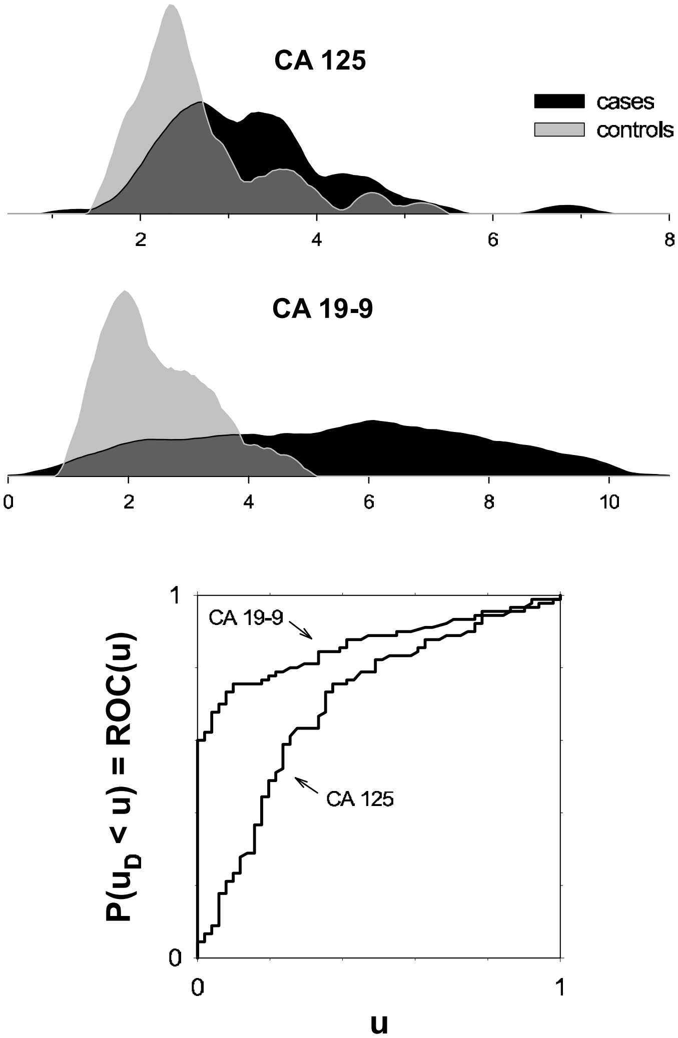

Figure 2: Two pancreatic cancer markers. Shown are frequency distributions in cases and controls (top two panels) and the associated placement value cdf's (lower panel).

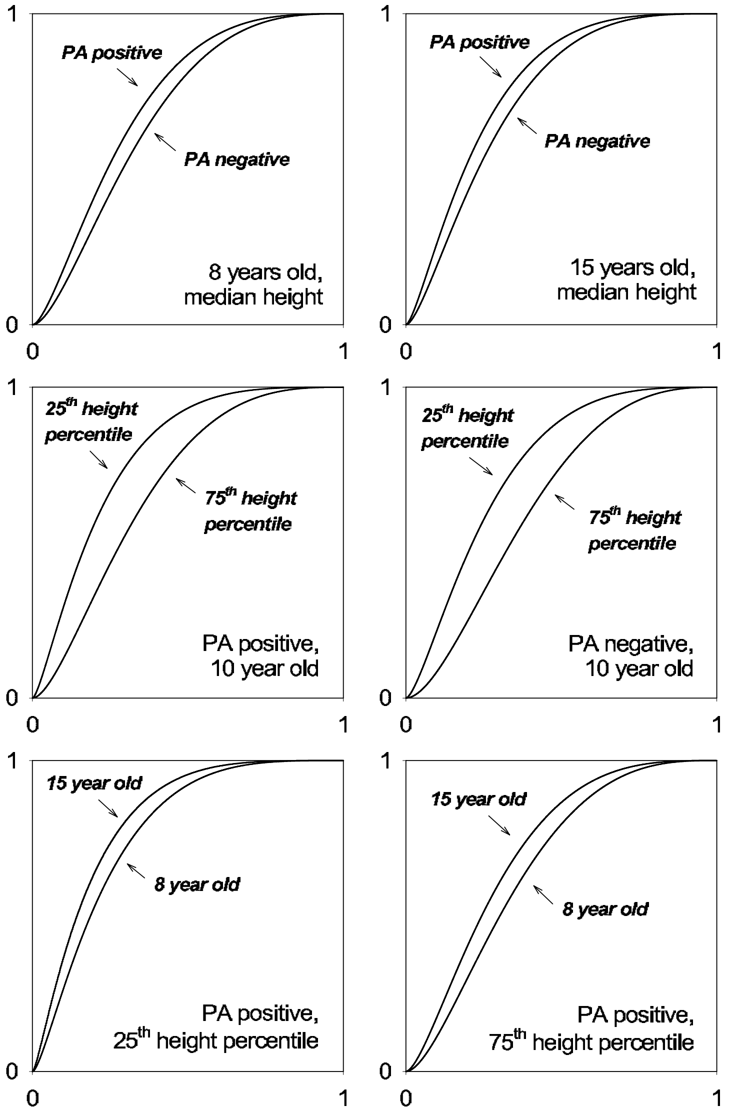

Figure 3: The placement value cdf's for FEV_1 in cystic fibrosis patients.