Biostatistics (2000), **1**, 2, *pp*. 123–140
Printed in Great Britain

Combining diagnostic test results to increase accuracy

MARGARET SULLIVAN PEPE

Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, PO Box 19024, Seattle, WA 98109-1024, USA

MARY LOU THOMPSON

Department of Biostatistics, University of Washington Seattle, WA 98195, USA

SUMMARY

When multiple diagnostic tests are performed on an individual or multiple disease markers are available it may be possible to combine the information to diagnose disease. We consider how to choose linear combinations of markers in order to optimize diagnostic accuracy. The accuracy index to be maximized is the area or partial area under the receiver operating characteristic (ROC) curve.

We propose a distribution-free rank-based approach for optimizing the area under the ROC curve and compare it with logistic regression and with classic linear discriminant analysis (LDA). It has been shown that the latter method optimizes the area under the ROC curve when test results have a multivariate normal distribution for diseased and non-diseased populations. Simulation studies suggest that the proposed non-parametric method is efficient when data are multivariate normal.

The distribution-free method is generalized to a smooth distribution-free approach to: (i) accommodate some reasonable smoothness assumptions; (ii) incorporate covariate effects; and (iii) yield optimized partial areas under the ROC curve. This latter feature is particularly important since it allows one to focus on a region of the ROC curve which is of most relevance to clinical practice. Neither logistic regression nor LDA necessarily maximize partial areas. The approaches are illustrated on two cancer datasets, one involving serum antigen markers for pancreatic cancer and the other involving longitudinal prostate specific antigen data.

Keywords: Biomarkers; Classification; Disease screening; ROC curve; Sensitivity; Specificity.

1. INTRODUCTION

Diagnostic tests often yield more than one output parameter. Alternatively, several diagnostic tests may be performed simultaneously. A question which arises in such settings is how to combine information from multiple test results in order to discriminate diseased from non-diseased states. Let $Y_1, Y_2, \ldots, Y_P$ denote the distinct numeric test results. In this paper we consider linear combinations of test results, $S(\alpha, Y) = \Sigma\alpha_p Y_p$, and the choice of coefficients $(\alpha_1, \ldots, \alpha_P)$ which maximize the diagnostic accuracy associated with the resultant composite score, S.

Various measures of diagnostic accuracy might be used. Here we focus on the area under the ROC (receiver operating characteristic) curve as the objective function. This is the most widely used index of diagnostic accuracy for diagnostic tests with continuous or ordinal data (Begg, 1991). The ROC curve for a score, such as S, is defined as the set of points $\{(\text{FP}(c), \text{TP}(c)), c \in (-\infty, \infty)\}$ where $\text{TP}(c) = P[S_i > c|$ study unit i is diseased], which can be interpreted as the true positive rate associated with the positivity criterion $S > c$ and $\text{FP}(c) = P[S_j > c|$ study unit j is non-diseased], which can similarly

be interpreted as the false positive rate at threshold c. The ROC curve therefore shows the trade-offs between increasing true positive (TP(c)) and increasing false positive rates which are feasible with the diagnostic score. It is a monotone increasing function from (0, 0) to (1, 1), with curves closer to the (0, 1) point associated with better diagnostic tests. The area under the ROC curve is a summary measure of accuracy, lying in the range (0.5,1), with 1 indicating perfect discrimination for some threshold c and 0.5 indicating no discrimination capacity. Indeed it can be interpreted as $P(S_D > S_{\bar{D}})$ where S_D and $S_{\bar{D}}$ represent scores for randomly chosen diseased and non-diseased subjects. It measures the distance between the distributions of scores for diseased and non-diseased subjects, in a distribution-free sense. A related measure of diagnostic accuracy which we will also consider is the partial area under the ROC curve, which restricts attention to the ROC curve over a range of acceptable false positive rates for the test.

Other measures of diagnostic accuracy exist, such as predictive values. The positive predictive value (PPV) of a dichotomous test is P(disease|positive test) and correspondingly the negative predictive value (NPV) is P(not diseased|negative test). The natural analogues of PPV and NPV for the continuous score S are PPV(c) $= P$(disease|$S > c$) and NPV(c) $= P$(not diseased|$S < c$). These predictive value functions are defined relative to thresholds and since in practice clinical decisions will be based on the score exceeding a threshold, it is relevant to focus on measures of accuracy associated with such criteria. Unfortunately these have not yet been well studied in the literature. Thus we employ the better developed ROC approach here.

The objective then is to find α_{opt}, which is the $\{\alpha_1, \ldots, \alpha_P\}$ that maximizes the area under the ROC curve associated with $\sum \alpha_p Y_p$. It has been shown that if $\{Y_1, \ldots, Y_P\}$ has a multivariate normal distribution in each of the diseased and non-diseased populations, then the score defined by the linear discriminant function maximizes the area under the ROC curve (Su and Liu, 1993). In this paper we relax the multivariate normal assumption and seek linear forms which maximize a distribution-free estimate of the area under the curve. Interestingly this approach, in contrast to the normal theory discriminant approach, can be extended in two important ways. First, it can be extended to maximize the partial ROC area. Second, it can accommodate covariates which affect diagnostic accuracy, such as age, disease severity or timing of measurement relative to disease onset.

In Section 2 we describe the basic approaches to deriving linear scores. In Section 3 we extend the distribution-free approach in the two aforementioned directions. In Section 4, results of simulation studies are presented wherein finite sample properties of the methods are assessed. We conclude with a discussion in Section 5 of how our methods relate to other methods for deriving linear scores for classification.

2. THE BASIC APPROACHES

2.1. Preliminaries

We use a simple dataset to illustrate the ideas. These data were derived from a study of 90 pancreatic cancer patients and 51 control patients with pancreatitis (Wieand *et al.*, 1989). Two serum markers were measured on these patients, the cancer antigen CA125 and CA19-9 which is a carbohydrate antigen. For our purposes the marker values were transformed to a natural logarithmic scale and are displayed in Figure 1. Estimated ROC curves associated with ln(CA125), Y_2, and with ln(CA19-9), Y_1, are shown in Figure 2 with areas of 0.71 and 0.86, respectively. The objective in this analysis is to derive a linear combination of Y_1 and Y_2 which yields a better ROC curve than either one alone.

In this example, and throughout most of this paper, only two markers are involved. Dealing with the simplest setting, $P = 2$, we can address the fundamental issues, while avoiding some of the computational difficulties, which need to be addressed when $P > 2$. We will discuss extensions to settings with more than two markers later. With two markers, finding the linear combination $\alpha_1 Y_1 + \alpha_2 Y_2$ which maximizes

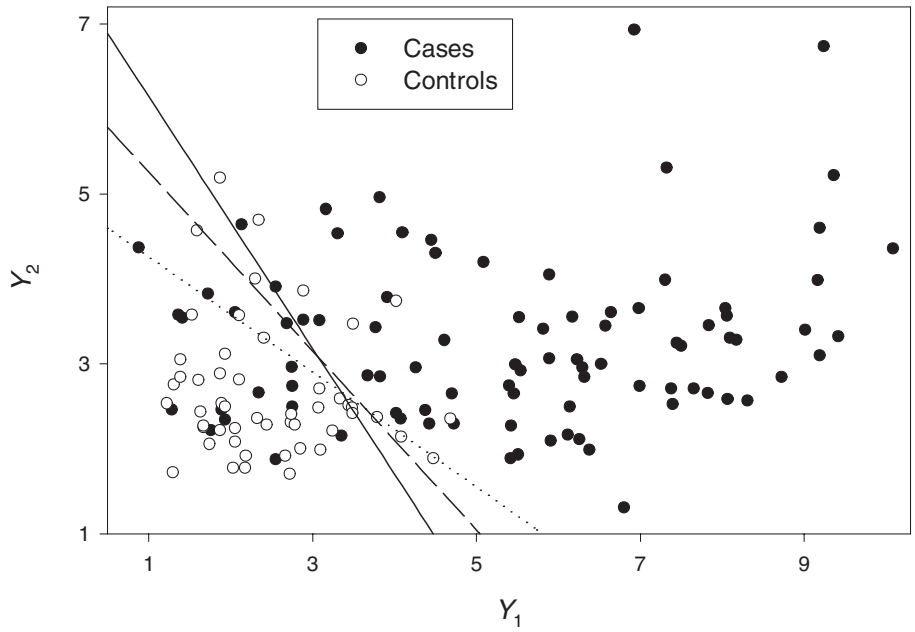

Fig. 1. $Y_2 = \ln(\text{CA125})$ versus $Y_1 = \ln(\text{CA19-9})$ for 141 subjects. Lines corresponding to estimated 80% specificity for scores calculated with the LD (—), DF (– – –), and LR (.....) methods are also shown.

the area under the ROC curve (AUC) is equivalent to finding the value $\alpha \in (-\infty, \infty)$ such that $Y_1 + \alpha Y_2$ maximizes the AUC. This is a consequence of the fact that the ROC curve is invariant to scale transformations. In each of our applications we standardize Y_1 and Y_2 to have mean 0 and variance 1, to assist in the interpretation of α as a relative weight of Y_2 to Y_1 in the combination.

2.2. The normal linear discriminant approach

If $Y = (Y_1, Y_2)'$ is distributed as a multivariate normal random variable with mean and variance-covariance parameters (μ^D, Σ^D) and $(\mu^{\bar{D}}, \Sigma^{\bar{D}})$, in the diseased and non-diseased population, respectively, then the AUC for $Y_1 + \alpha Y_2$ is

$$\text{AUC}_{\text{LD}}(\alpha) = \Phi\left((1,\alpha)(\mu^D - \mu^{\bar{D}}) / \left\{\sqrt{(1,\alpha)(\Sigma^D + \Sigma^{\bar{D}})(1,\alpha)}\right\}\right),$$

where Φ denotes the standard cumulative normal distribution function. As shown by Su and Liu (1993), the optimal coefficient α is $\alpha_{\text{LD}} = a_2/a_1$ where

$$(a_1, a_2)' = \left(\Sigma^D + \Sigma^{\bar{D}}\right)^{-1} (\mu^D - \mu^{\bar{D}}).$$

Moreover, the optimal AUC is

$$\text{AUC}_{\text{LD}}^{opt} = \Phi\left(\sqrt{\left\{(\mu^D - \mu^{\bar{D}})' \left(\Sigma^D + \Sigma^{\bar{D}}\right)^{-1} (\mu^D - \mu^{\bar{D}})\right\}}\right).$$

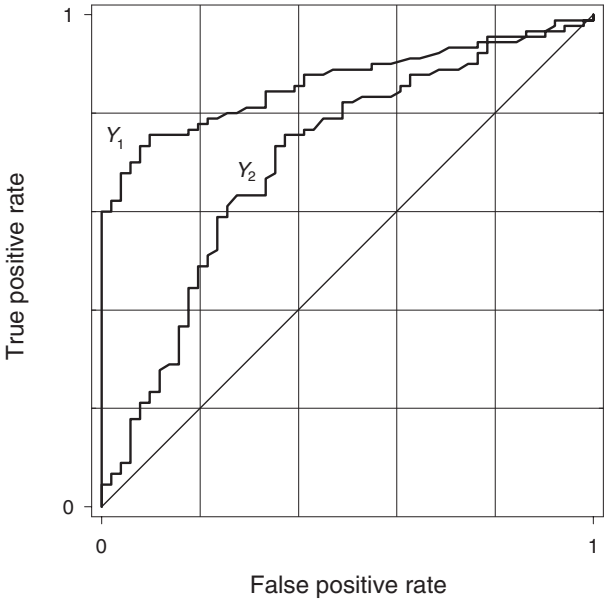

Fig. 2. ROC curves for $Y_1 = \ln(\text{CA19-9})$ and $Y_2 = \ln(\text{CA125})$ based on data from 90 cases and 51 controls. The areas under the ROC curves are 0.86 for Y_1 and 0.71 for Y_2.

In practice the mean and variance-covariance parameters are estimated from the data and the estimates are substituted into the above formulas.

Applying these results to the pancreatic cancer data yielded $\hat{\alpha}_{\text{LD}} = 0.27$ and $\widehat{\text{AUC}}_{\text{LD}}^{\text{opt}} = 0.893$. Thus, a linear combination of (standardized) Y_1 and Y_2 with Y_1 receiving higher weight than Y_2 appears to have the best discriminating capacity among all linear combinations of Y_1 and Y_2. Note that, the analysis does not suggest a specific threshold or decision rule be associated with the combination. Such optimization would require information on costs associated with false positive and false negative errors as well as information on disease prevalence. In the absence of such information, which may vary with the application, our results simply provide us with the optimal linear combination that maximizes the AUC distance between the diseased and non-diseased populations. One might use different thresholds for the combination in different applications. Figure 1 shows the 'threshold' line corresponding to (standardized) $Y_1 + \hat{\alpha}_{\text{LD}} Y_2 = c$, for estimated specificity of 80%. Parallel shifts of this line correspond to different choices of threshold (and hence different sensitivity and specificity) for the linear combination of the markers.

Figure 3 displays $\widehat{\text{AUC}}_{\text{LD}}(\alpha)$ versus α for $\alpha \in (-\infty, \infty)$. For ease of presentation, the plot displays $\widehat{\text{AUC}}(\alpha)$ versus $1/\alpha$ when $\alpha > 1$ or $\alpha < -1$. In this plot observe that $\widehat{\text{AUC}}(\alpha)$ corresponds to the AUC associated with Y_1 alone at $\alpha = 0$ and to that for Y_2 alone when $1/\alpha = 0$. Interestingly, the AUC for Y_1 alone appears to be similar to that for the optimal linear combination, $S = Y_1 + 0.27\, Y_2$ suggesting that Y_2, in fact, adds little to the discriminating capacity of Y_1.

2.3. The distribution-free approach

The above calculations pertain to the setting where (Y_1, Y_2) are assumed to have bivariate normal distributions. We now consider maximizing the AUC without assumptions on the distributions of (Y_1, Y_2). It has been shown (Bamber, 1975) that the area under the ROC curve for S can be interpreted as a prob-

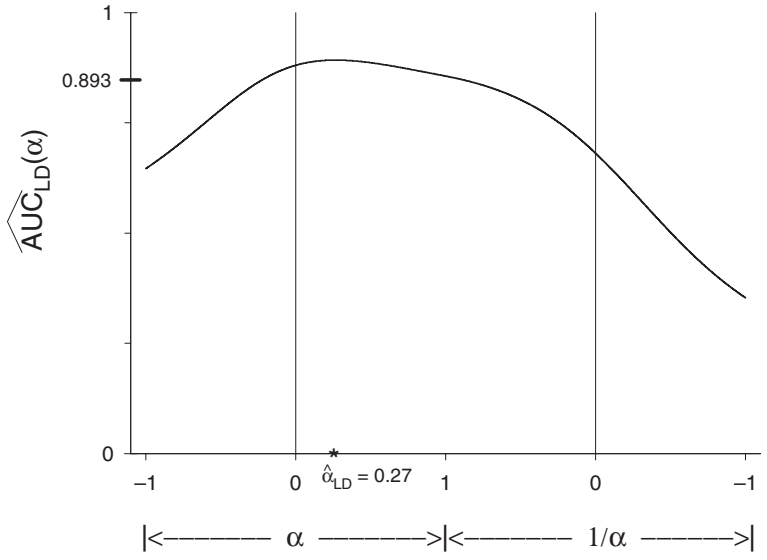

Fig. 3. Plot of the estimated area under the ROC curve associated with $Y_1 + \alpha Y_2$ when Y_1 and Y_2 are assumed to have a bivariate normal distribution and the LD method is applied to the pancreatic cancer marker data. The optimal coefficient is indicated with an asterisk.

ability, $P[S^D \geq S^{\bar{D}}]$, where S^D and $S^{\bar{D}}$ are scores for independent, randomly selected study units from the diseased and non-diseased populations, respectively. A rank-based estimate of the AUC, based on this fact, is the Mann–Whitney U statistic (Hanley and McNeil, 1982). If the data for diseased study units are denoted by $\{(Y_{i1}^D, Y_{i2}^D)\ i = 1, \ldots, n^D\}$ and that for non-diseased units are $\{(Y_{j1}^{\bar{D}}, Y_{j2}^{\bar{D}})\ j = 1, \ldots, n^{\bar{D}}\}$, the Mann–Whitney U-statistic estimator of the AUC associated with $S(\alpha, Y) = Y_1 + \alpha\ Y_2$ is

$$\widehat{\text{AUC}}_{\text{DF}}(\alpha) = \sum_{i=1}^{n^D} \sum_{j=1}^{n^{\bar{D}}} I\left[Y_{i1}^D + \alpha Y_{i2}^D \geq Y_{j1}^{\bar{D}} + \alpha Y_{j2}^{\bar{D}}\right] \Big/ n^D n^{\bar{D}},$$

where the subscript indicates that it is a distribution-free (DF) estimator. Therefore, as an estimate of the optimal coefficient, α_{opt}, one might choose the α that maximizes the Mann–Whitney U statistic and denote it by $\hat{\alpha}_{\text{DF}}$. Since $\widehat{\text{AUC}}_{\text{DF}}(\alpha)$ is not a continuous function of α, a search rather than a derivative-based method is required for this maximization.

To implement the maximization on the pancreatic cancer data, we evaluated $\widehat{\text{AUC}}_{\text{DF}}(\alpha)$ for 201 equally spaced values of α in $[-1,1]$. For $\alpha < -1$ and $\alpha > 1$ we note that the AUC for $Y_1 + \alpha Y_2$ is the same as that for $\gamma Y_1 + Y_2$ where $\gamma = 1/\alpha \in [-1, 1]$. Thus we also evaluated the AUCs pertaining to $\gamma Y_1 + Y_2$ for 201 equally spaced values of γ in $[-1, 1]$. The procedure is therefore symmetric in its treatment of Y_1 and Y_2. The optimal coefficient for Y_2 is the α in $[-1,1]$ or the γ^{-1} where $\gamma \in [-1, 1]$ which maximizes the $\widehat{\text{AUC}}_{\text{DF}}$.

The optimal weighting for Y_2 using this non-parametric procedure applied to the prostate cancer was estimated as $\hat{\alpha}_{\text{DF}} = 0.39$ with associated $\widehat{\text{AUC}}_{\text{DF}}$ being 0.894. This is very close to the optimized area of 0.893 found using the normal theory approach. Indeed, comparing Figures 3 and 4, estimated areas $\widehat{\text{AUC}}_{\text{LD}}(\alpha)$ and $\widehat{\text{AUC}}_{\text{DF}}(\alpha)$ were very similar for these data for all values of α.

Because the distribution-free (DF) approach does not depend on assumptions about the joint distri-

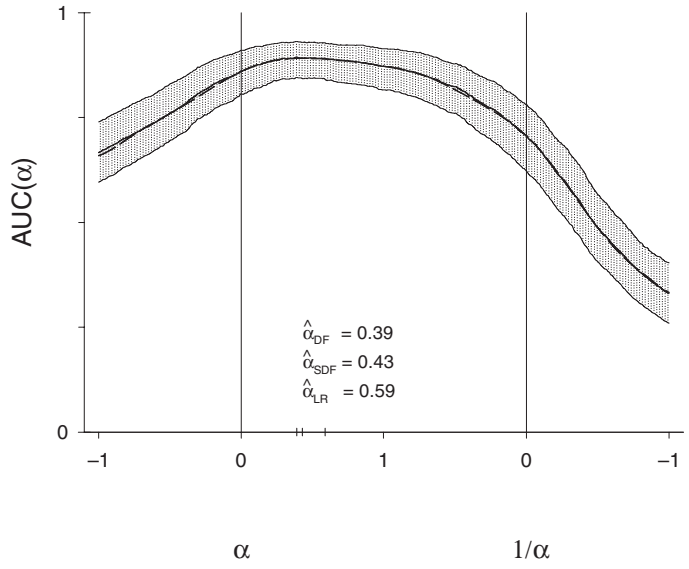

Fig. 4. AUC(α) for the pancreatic cancer data estimated using the distribution-free method (solid curve) along with the smooth distribution-free method (broken curve). Also shown are pointwise 90% confidence intervals based on the distribution-free estimator and its bootstrap distribution.

bution of (Y_1, Y_2), but the binormal linear discriminant analysis method does, we would expect the DF procedure to be more robust. To illustrate this, admittedly in an extreme case, Figure 5(a) displays simulated data for which the linear score $Y_1 + Y_2$ perfectly discriminates diseased from non-diseased states. The DF procedure determined the true value of α_{opt}, i.e. $\hat{\alpha}_{\text{DF}} = 1$ with these data. With the linear discriminant procedure, however, $\hat{\alpha}_{\text{LD}} = -0.106$ which does not yield optimal discrimination. The empirically estimated ROC curves associated with the two scores, $Y_1 + \hat{\alpha}_{\text{DF}} Y_2$ and $Y_1 + \hat{\alpha}_{\text{LD}} Y_2$, shown in Figure 5(b), have empirically estimated AUC statistics of 1.00 and 0.91, respectively.

2.4. Logistic regression

Logistic regression has been proposed as a means of modelling the probability of disease given several test results (Richards *et al.*, 1996). It yields a linear score that intuitively discriminates diseased from non-diseased subjects. It is well known that in the multivariate binormal setting, when the distributions of $(Y_1, \ldots, Y_P)$ are multivariate normal in the diseased and non-diseased populations, the linear discriminant and logistic scores are equal if the variance-covariance matrices are proportional. The linear discriminant procedure has been shown to be statistically more efficient (Efron, 1975) when the model is correct. Logistic regression, however, can be applied outside of the multivariate binormal setting. It relies only on an assumption about the form of the conditional probability for disease given $(Y_1, \ldots, Y_P)$ and does not require specification of the much more complex joint distribution of $(Y_1, \ldots, Y_P)$.

We consider logistic regression here as an alternative means to derive a linear score because the procedure is widely available and easy to use. In contrast to the LDA and DF procedures, however, it is not motivated as a procedure which maximizes the area under the ROC curve for the linear score. Rather, in logistic regression analysis, the coefficients $(\alpha_1, \ldots, \alpha_P)$ are chosen to maximize the logistic likelihood. It is not clear if the logistic likelihood relates to any natural measure of the discriminatory capacity of the

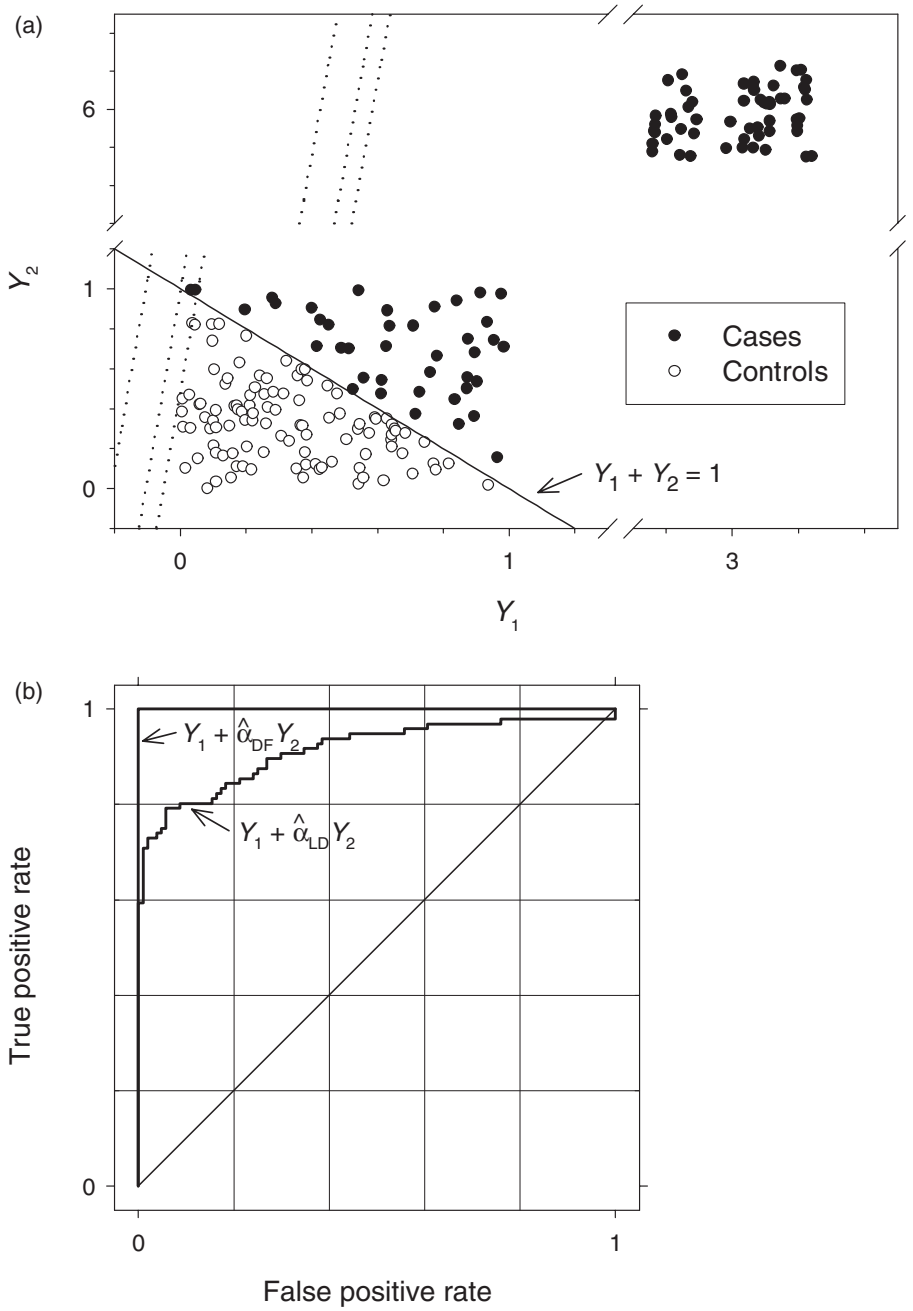

Fig. 5. (a) Hypothetical data for which $Y_1 + Y_2 > 1$ for cases and $Y_1 + Y_2 < 1$ for controls. Dotted lines indicate $Y_1 + \hat{\alpha}_{\mathrm{LD}} Y_2 = k$ for various values of k and show poor discrimination between cases and controls. (b) Empirical ROC for the linear combination $Y_1 + \alpha_{\mathrm{LD}} Y_2$ based on hypothetical data shown in (a). The AUC is 0.91. The ROC for the score $Y_1 + \alpha_{\mathrm{DF}} Y_2 = Y_1 + Y_2$ is perfect, with sensitivity = 1, specificity = 1, and AUC = 1.

linear score. Hence, in general, the logistic regression linear score is not easily motivated as an optimal discriminator of diseased and non-diseased populations except in the multivariate binormal setting. It has been shown, however, that, if complete discrimination is possible, LR will estimate the linear combination which separates the populations (Day and Kerridge, 1967). This characteristic is illustrated below.

We applied the logistic regression model, $P(D = 1 \mid Y_1, Y_2) = \text{logit}^{-1}(\beta_0 + \beta_1 Y_1 + \beta_2 Y_2)$, to the pancreatic cancer data. The estimated logistic coefficients were $\beta_1 = 0.027$ for Y_1 and $\beta_2 = 0.016$ for Y_2 thus yielding $\hat{\alpha}_{\text{LR}} = 0.016/0.027 = 0.594$ for the corresponding estimate of α_{opt}. Since the logistic likelihood depends on each of the regression coefficients, we cannot plot the likelihood, i.e. the objective function, as a simple function of $\alpha = \beta_2/\beta_1$. Thus, there is no simple plot for logistic regression that corresponds to the plots in Figures 3 and 4 for the linear discriminant and distribution-free procedures. The empirical estimate of the AUC associated with $\hat{\alpha}_{\text{LR}} = 0.594$ is 0.891 and is indicated in Figure 4.

When logistic regression was applied to the data in Figure 5(a), the estimated coefficients were extremely large with values of 797 for β_1, 806 for β_2 and -801 for the constant term β_0, essentially estimating a logistic probability function which transitions rapidly from 0 to 1 where $Y_1 + \alpha_{\text{LR}} Y_2 = 1$. Thus $\hat{\alpha}_{\text{LR}} = 1.01$, which is almost exactly the true value of 1.00 as anticipated.

2.5. A smooth distribution-free approach

Our next approach is a modification of the distribution-free method. Like the DF approach, it is based on the fact that $\text{AUC}(\alpha) = P[Y_1^D + \alpha Y_2^D \geq Y_1^{\bar{D}} + \alpha Y_2^{\bar{D}}]$ but incorporates the assumption that $\text{AUC}(\alpha)$ is a smooth function of α when Y_1 and Y_2 are continuous random variables. It will be shown that it generalizes the DF approach and provides important additional capabilities that are not offered by discriminant analysis or logistic regression. The idea is to model $\text{AUC}(\alpha)$ as a smooth function of α:

$$\text{logit}\{\text{AUC}(\alpha)\} = \beta(\alpha),$$

where $\beta(\alpha)$ is say a cubic regression spline, and to use a device based on pairs of diseased and non-diseased observations to fit the model.

For each of the $n^D \times n^{\bar{D}}$ combinations $(Y_{i1}^D, Y_{i2}^D, Y_{j1}^{\bar{D}}, Y_{j2}^{\bar{D}})$ of a diseased and non-diseased observation $\{i = 1, 2, \ldots, n^D;\ j = 1, 2, \ldots, n^{\bar{D}}\}$, choose a set of m possible weightings $\{\alpha_1^{ij}, \alpha_2^{ij}, \ldots, \alpha_m^{ij}\}$ for Y_2 and define

$$U_{ijk} = I\left[Y_{i1}^D + \alpha_k^{ij}\, Y_{i2}^D > Y_{j1}^{\bar{D}} + \alpha_k^{ij}\, Y_{j2}^{\bar{D}}\right]$$

for $k = 1, \ldots, m$. Recall that $E\{U_{ijk}\} = \text{AUC}(\alpha_k^{ij})$. Moreover, we propose to model $\text{AUC}(\alpha_k^{ij})$ as a regression spline in α_k^{ij}, i.e. as a linear combination of the regression spline basis functions of α_k^{ij}. We, therefore, create the basis functions of α_k^{ij} as covariates for U_{ijk} and fit the model for $\text{AUC}(\alpha_k^{ij})$ using standard GLM binary regression methods. Details of the fitting procedure can be found in the Appendix, including choice of weightings, $\{\alpha_k^{ij}\}$, to ensure that the curve is fit over the entire domain $\alpha \in (-\infty, \infty)$. A simple search applied to the fitted curve, $\widehat{\text{AUC}}_{\text{SDF}}(\alpha)$, yields the $\hat{\alpha}_{\text{SDF}}$ which maximizes the AUC.

Results of such a fitting procedure applied to the pancreatic cancer data are shown in Figure 4. In this analysis we chose $m = 40$ and weights $\{\alpha_1^{ij}, \ldots, \alpha_m^{ij}\}$ chosen at random for each (i, j), with half having a uniform distribution in $(-1, 1)$ and half such that $\gamma = \alpha^{-1}$ had a uniform distribution in $(-1, 1)$. The smooth curve approximates the AUC curves generated by LDA and the DF methods closely (Figures 3 and 4). The optimal relative weighting of Y_1 and Y_2 corresponds to $\hat{\alpha}_{\text{SDF}} = 0.43$ with estimated area of $\widehat{\text{AUC}}_{\text{SDF}} = 0.891$.

Interestingly, the distribution-free method described in Section 2.3 can be derived as a special case of the smooth distribution-free binary regression-based method. If $\{\alpha_1^{ij}, \ldots, \alpha_m^{ij}\}$ are chosen to be the same

for all (i, j) pairs and if the logistic model is saturated in the sense that it includes a distinct parameter for each α (rather than using a regression spline basis in α) then, using the logistic approach, the fitted value at each α_k will be the proportion of U_{ijk} equal to 1. This is exactly the Mann–Whitney U statistic, $\widehat{\mathrm{AUC}}_{\mathrm{DF}}(\alpha_k)$. Since the fitted values in this case are the same as for the DF procedure, the maximizing α will be the same. In general, the smooth distribution-free (SDF) method seeks to estimate $\mathrm{AUC}(\alpha)$ with some smoothness constraints not imposed by the simple DF method. This may lead to increased statistical efficiency. Our main interest in the SDF method, however, is in the important extensions which can be achieved, which we describe next.

3. Extensions of SNP

3.1. Optimizing partial areas

As an alternative to the area under the entire ROC curve as a measure of accuracy, Thompson and Zucchini (1989) and McClish (1989) have suggested using the area under the ROC curve in a restricted domain of false positive rates, the so-called partial area under the ROC curve (pAUC). The rationale is that in many applications, tests with false positive rates outside of a particular domain will be of no practical use and hence are irrelevant for evaluating the accuracy of the test. For instance, for a condition with low prevalence, it may be clear that the referrals resulting from high false positive rates will put an unacceptable burden on the health care system. Thus restricting attention to the ROC curve over a practically relevant range of false positive rates is appealing. Wieand *et al.* (1989) also argue this point.

Let $[0, t_0]$ denote the range of false positive rates of potential interest and as before let Y^D and $Y^{\bar{D}}$ denote random independent observations from diseased and non-diseased sets. It follows from Pepe (1997) that

$$\mathrm{pAUC}(\alpha) = P\left[Y_1^D + \alpha Y_2^D > Y_1^{\bar{D}} + \alpha Y_2^{\bar{D}} \text{ and } Y_1^{\bar{D}} + \alpha Y_2^{\bar{D}} > Q^{\bar{D}}(1 - t_0, \alpha)\right],$$

where $Q^{\bar{D}}(1 - t_0, \alpha)$ is the $(1 - t_0)$ quantile of $Y_1^{\bar{D}} + \alpha Y_2^{\bar{D}}$. We can use this result to optimize the partial area using the SDF approach. The idea is to estimate the $(1 - t_0)$ quantiles of $Y_1^{\bar{D}} + \alpha Y_2^{\bar{D}}$ using data from the non-diseased set of observations, as is described in the Appendix. Then the SDF method is applied as described in Section 2.5, except that U_{ijk} is set to 0 if $Y_{j1}^{\bar{D}} + \alpha_k^{ij} Y_{j2}^{\bar{D}} < \hat{Q}^{\bar{D}}(1 - t_0, \alpha_k^{ij})$.

3.2. Incorporating covariates

Characteristics of study subjects or features of the measurement process can influence the performance of a diagnostic score and hence the area under its ROC curve. If such covariates are available and denoted by X one can account for them in the SDF analysis by including them in a logistic regression model, such as:

$$\mathrm{logit}\{\mathrm{AUC}(\alpha, X)\} = \beta(\alpha) + \tau X.$$

This particular model assumes that $\beta(\alpha)$ is the same for all X.

The optimal linear combination $Y_1 + \alpha Y_2$ can be found with the SDF procedure in the following way. Let X_{ij} denote the covariates pertinent to the ith diseased and jth non-diseased observations. In fitting the logistic model to the binary indicator variables $\{U_{ijk}, i = 1, \ldots, n^D, j = 1, \ldots, n^{\bar{D}}, k = 1, \ldots, m\}$, in addition to the regression spline basis functions of α_k^{ij}, the covariates X_{ij} are also included as predictor variables. The optimal α is then identified by a simple search to maximize the fitted regression spline $\hat{\beta}(\alpha)$. Observe that with the above model this approach results in a single linear combination which,

assuming that the model is correct, is optimal for all covariate values. Indeed, since the same α optimizes $\mathrm{AUC}(\alpha, X)$ for all X, it will also optimize the marginal area function $\mathrm{AUC}(\alpha) = E(\mathrm{AUC}(\alpha, X))$. In large samples an analysis that ignores X will therefore yield the same optimal α. Efficiency gains, however, would be expected from fitting the covariate adjusted model.

For certain types of covariates it might be of interest to derive optimized scores which can vary with covariate values, i.e. $S(\alpha, Y, X) = Y_1 + \alpha(X)Y_2$. This is easily accomplished by stratification when X is categorical. More generally, however, one can use a parametric function for $\alpha(X)$. For example, one could fit a model

$$\text{logit } \{\mathrm{AUC}(\alpha, X)\} = \beta_1(\alpha) + \tau\ X + \beta_2(\alpha)X.$$

This allows the optimal, α, to vary with X. For any particular X, the optimal α, $\alpha(X)$, is that which maximizes $\beta_1(\alpha) + \beta_2(\alpha)X$.

It should be noted that in many applications it will be desirable to derive a single combination of markers which is not covariate specific. In particular this is true for covariates with values which cannot be known at the time of testing in practical applications. Examples of such covariates are disease severity (Pepe, 1998) and timing of test relative to clinical diagnosis (Etzioni *et al*., 1999).

3.3. Illustrations with prostate cancer data

We now illustrate the extended SDF approach on a prostate cancer dataset which is more complex than the pancreatic dataset. Prostate specific antigen (PSA) measured in serum is currently used as a biomarker for prostate cancer. To better understand its potential role in screening, free and bound levels of PSA were measured in sera for 71 subjects who developed prostate cancer and for 71 age-matched controls, all of whom participated in the CARET study, a randomized lung cancer prevention study including 12 025 men (Thornquist *et al*., 1993). Subjects who participated in CARET had serum drawn and stored at entry into the study and at 2 year intervals thereafter. Blood samples drawn after diagnosis of prostate cancer were excluded from this analysis, leaving on average 3.2 samples per case and 6.5 samples per control in the dataset (Figure 6).

Two different measures of PSA have been proposed in the literature for screening, total PSA and the ratio of free to total PSA. Let $Y_1 = \log(\text{total PSA})$ and $Y_2 = -\log(\text{free PSA/total PSA})$, both of which tend to be larger in cases than in controls. Etzioni *et al*. (1999) and Pearson *et al*. (1996) have compared the diagnostic values of Y_1 and Y_2, each used on its own. Here we consider how Y_1 and Y_2 might be used together. Although equivalently we could consider $Y_2 = -\log(\text{free PSA})$, we chose $Y_2 = -\log(\text{free PSA/total PSA})$ because it is a measure of interest in itself which has been used in the literature and we will want to compare the accuracy associated with it to that of the linear combination. A covariate which affects diagnostic accuracy in this setting concerns the timing of the serum sample relative to clinical diagnosis of disease for cases. Let T_i be the time prior to diagnosis at which the serum sample i is drawn for a case. Accuracy would be expected to increase with decreasing values of T_i.

We fit the model $\text{logit } \{\mathrm{AUC}(\alpha, T)\} = \beta(\alpha) + \tau T$ to the data using the SDF approach. An interaction between α and T was not considered because T is a variable which will not be known at the time of screening in future applications. Thus a single linear combination which does not depend on T is desired. Although cases and controls contributed several serum samples to the analysis, the data analysis unit pertained to serum sample, rather than case or control per se. Due to the large number of observations involved, in order not to exceed the storage capacity of our computer, we chose $m = 2$. The fitted model is displayed in Figure 7. The optimal linear combination is $Y_1 + 0.37\ Y_2$ with optimized area being 0.894 at $T = 0$ and 0.835 at $T = 4$ years. The optimized area was similar to that obtained for Y_1 alone as a marker ($\mathrm{AUC}(0, 0) = 0.889$; $\mathrm{AUC}(0, 4) = 0.828$) and thus in this dataset it does not appear that the ratio measure adds substantially to accuracy when used in linear combination with total PSA. Interestingly,

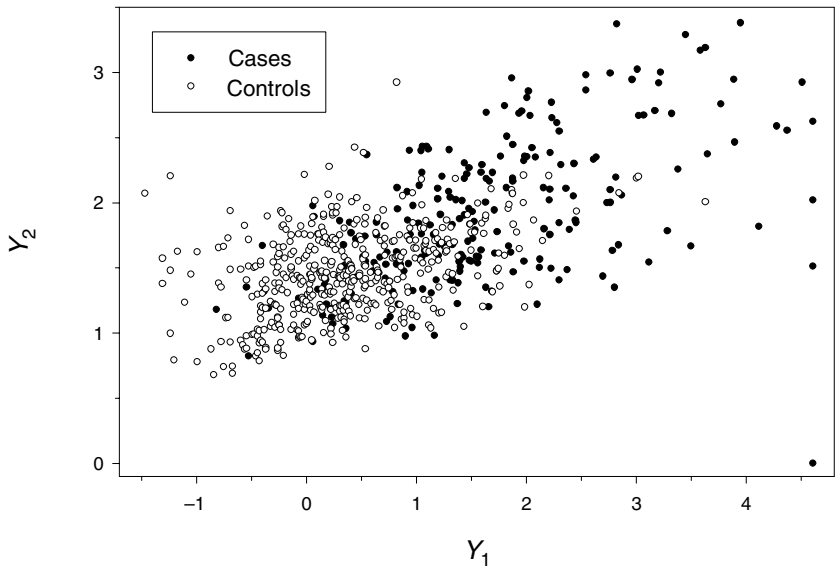

Fig. 6. Two measures of PSA, $Y_1 = \ln$(total serum PSA) and $Y_2 = -\ln$(free PSA/total serum PSA) measured on 71 prostrate cancer cases and 71 controls. Measurements were taken serially in time.

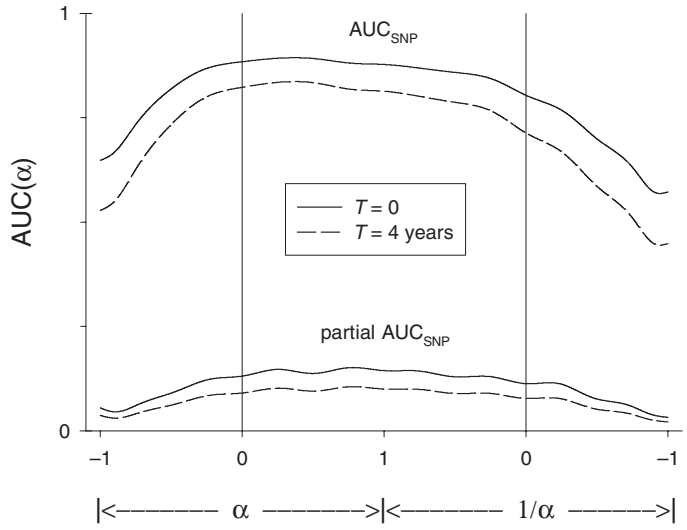

Fig. 7. Smooth non-parametric AUC and partial AUC estimates for prostate cancer data markers based on the model logit(AUC(α, T)) $= \beta(\alpha) + \gamma T$ where T is time of measurement relative to diagnosis for cases.

relative to using either marker alone it appears that accuracy is substantially reduced when the difference between marker values is considered. This can be seen by the fact that the AUC decreases for negative values of α. A likely explanation for this is the positive correlation between the markers (Figure 6). In this example, the optimal coefficient $\hat{\alpha}_{\text{SDF}}$ was robust to the manner in which T entered the model, taking the value 0.36 when $-T + 1$ was transformed to a logarithmic scale.

A model which excludes the covariate T was fit to the same data, logit $(\mathrm{AUC}(\alpha)) = \beta(\alpha)$. This yielded the same value of $\alpha = 0.37$ to maximize the area under the curve. Observe that the optimized area in this model is not specific to T but rather is, in a sense, averaged over observed T. We found $\hat{\mathrm{AUC}}(\alpha) = 0.850$. A resampling experiment was done to ascertain the variability in the optimal α when α was derived using the models with and without adjustment for the covariate T. Based on 100 resampled datasets we found $\mathrm{var}(\hat{\alpha}_{\mathrm{opt}}) = 0.064$ without covariate adjustment and $\mathrm{var}(\hat{\alpha}_{\mathrm{opt}}) = 0.057$ with covariate adjustment, a gain of 12% in efficiency by including T in the model.

When T is ignored as a factor relating to the predictive capacity of the marker, the LDA and LR procedures can be applied as described earlier. They yield $\hat{\alpha}_{\mathrm{LDA}} = 0.276$ and $\hat{\alpha}_{\mathrm{LR}} = 0.325$, respectively. Note that neither of these approaches can accommodate a covariate such as T which is specific to diseased observations. In such settings only the SDF approach appears to be an option at present.

We next implemented the SDF approach to optimize the partial AUC. Again, neither LDA nor LR can target the partial AUC specifically as the SDF method can. Restricting attention to false positive rates $\leq 20\% = t_0$ with the SDF approach we fit the model logit $\mathrm{pAUC}(\alpha, T) = \beta(\alpha) + \tau T$ to the data. The same choices for m and for the regression spline knots (see Appendix) were used as above. The fitted model is also displayed in Figure 7. The maximized partial AUC is achieved with $\alpha = 0.78$ which is substantially different than the relative weighting of $\alpha = 0.37$ found when optimizing the full AUC. The procedure yielded an estimated optimal pAUC of 0.106 at 4 years prior to diagnosis and 0.152 at the time of clinical diagnosis. To interpret these values, consider that the maximum achievable pAUC is 0.20, and the pAUC for an uninformative test is $(0.20)^2/2 = 0.02$. Thus, the linear combination, $Y_1 + 0.78Y_2$, appears to yield an informative score in the sense that it has a good ROC curve over an important range of false positive rates. Note that although, in this example, the linear combination which maximizes the partial area differs from that which maximizes the total area, both area measures are relatively flat functions of α.

4. EFFICIENCY

Though, as demonstrated in Section 2.3, the distribution-free methods are more robust than the linear discriminant procedure, presumably they incur some loss of efficiency relative to linear discriminant analysis (LDA) when the test result data follow a bivariate normal distribution. Simulation studies were conducted therefore to investigate the extent of this loss in efficiency. We focused on the comparison of LDA with the fully non-parametric or (DF) approach, reasoning that this provides a picture of the most extreme efficiency loss and that because of smoothing the performance of the SDF approach would be intermediate between the LDA and DF methods.

Bivariate normal data, (Y_1, Y_2), were generated for n^D cases and $n^{\bar{D}}$ controls with mean and variance-covariance matrix for cases

$$\mu^D = \begin{pmatrix} \delta_1 \\ \delta_2 \end{pmatrix}, \qquad \Sigma^D = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix},$$

and for controls

$$\mu^{\bar{D}} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \qquad \Sigma^{\bar{D}} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Without loss of generality $\delta_1 > \delta_2 > 0$ and we considered $\rho \geq 0$ to be of most practical interest. Thus the ROC curve for Y_p alone $(p = 1, 2)$ is equal to $\left\{\mathrm{ROC}(f) = \Phi\left(\delta_p + \Phi^{-1}(f)\right);\ f \in (0, 1)\right\}$ with $\mathrm{AUC} = \Phi(\delta_p/\sqrt{2})$. The ROC curve associated with $Y_1 + \alpha Y_2$ has area:

$$\mathrm{AUC}(\alpha) = \Phi\left((\delta_1 + \alpha\delta_2)/\sqrt{(2\{1 + 2\alpha\rho + \alpha^2\})}\right).$$

This, according to Su and Liu (1993), is optimized at

$$\alpha_{\text{opt}} = (1 - \rho\delta_1/\delta_2) \Big/ (\delta_1/\delta_2 - \rho)\,.$$

When Y_1 and Y_2 are equally accurate on their own, i.e. $\delta_1 = \delta_2$, the optimal linear combination is $Y_1 + Y_2$. Otherwise, the more accurate test (Y_1) is given more weight in the optimal linear combination, i.e. $\alpha_{\text{opt}} < 1$ if $\delta_1 > \delta_2$.

Table 1 shows the results of 1000 simulations, in each of which the LDA and DF methods were applied to the same data. In addition to summaries of derived coefficients, $\hat{\alpha}_{\text{DF}}$ and $\hat{\alpha}_{\text{LDA}}$, and optimal estimated areas, $\widehat{\text{AUC}}_{\text{DF}}(\hat{\alpha}_{\text{DF}})$ and $\widehat{\text{AUC}}_{\text{LDA}}(\hat{\alpha}_{\text{LDA}})$, we show measures of the true accuracies associated with the scores $Y_1 + \hat{\alpha}_{\text{DF}}Y_2$ and $Y_1 + \hat{\alpha}_{\text{LDA}}Y_2$. These are the true AUCs and points on the true ROC curve. Specifically we show the true positive rates for the optimized scores corresponding to the false positive rates of 0.05, 0.10 and 0.20 for them, which are denoted by TP(FP = 0.05), TP(FP = 0.10) and TP(FP = 0.20), respectively, in the tables.

Observe that the combination of Y_2 with Y_1 provides substantially better discrimination than does Y_1 alone, i.e. $\alpha = 0$, and that information on Y_2 is most beneficial when Y_2 is uncorrelated with Y_1. With an optimal linear combination the AUC is increased from 0.80 for Y_1 alone to 0.88 when $\rho = 0$ and to 0.83 when $\rho = 0.50$. On the ROC scale itself, at a false positive rate of 0.20, the disease detection rate increases from 0.64 to 0.80 when $\rho = 0$ and to 0.70 when $\rho = 0.50$.

The median coefficients α chosen by both the LDA and DF methods were approximately at the true optimal value, $\alpha = 1.00$. Variability in the coefficients was greater for the DF than LDA methods. On the more relevant scale of accuracy, however, differences between the methods seemed minor. For example, at $n^D = n^{\bar{D}} = 100$ and $\rho = 0$, the true median AUCs were the same (0.882) with 10–90th percentile ranges that differed by 0.002. Relative to the gain in accuracy achieved by including Y_2 with Y_1, the extra variation in AUC($\hat{\alpha}_{\text{DF}}$) over AUC($\hat{\alpha}_{\text{LD}}$) seems minor. Even with smaller sample sizes, $n^D = n^{\bar{D}} = 50$ and when Y_2 is less informative ($\rho = 0.50$) the difference between the methods is small. Consider that at a specificity of 0.90, the sensitivity increases optimally from 0.325 to 0.393. Both methods yield a median sensitivity of 0.390 with the difference between the lower 10th percentiles being only $0.372 - 0.364 = 0.008$. We conclude that with bivariate binormal data the rank-based DF procedure yields a linear combination with accuracy close to that of the optimal linear discriminant procedure in the settings we have studied. As suggested by a referee, we also applied the logistic regression (LR) procedure to the simulated data (results not shown). Not surprisingly the performance of the LR estimated score, $Y_1 + \hat{\alpha}_{\text{LR}}Y_2$, was intermediate between the LDA and DF scores. For binormal data with equal variance-covariance matrices, logistic regression yields consistent estimates of the linear discriminant function, that are less efficient than the maximum likelihood estimates produced by LDA. Because they impose some structure on the data, however, one would expect slightly better efficiency than the DF estimator. Given the close performance of the DF and LDA methods with binormal data, the gains by LR over DF procedures are not likely to be practically important.

5. Discussion

Statistical approaches to classification with multiple markers abound. These include binary regression, linear and non-linear discriminant analysis, decision trees, Bayesian decision making and neural networks. The latter three schemes derive specific decision rules to optimize an objective function. In contrast, our methodology derives a score but not a specific decision rule. Information on costs associated with errors and information on disease prevalence, for example, would be necessary in order to sensibly derive a specific decision rule based on the score. In the absence of such information we derive only a score.

Table 1. *Results of 1000 simulations of bivariate normal data. The AUC for Y_1 and Y_2 is* 0.80*, i.e.* $\delta = 1.19$*. Shown are medians with 10th and 90th percentiles in parentheses*

	$\rho = 0.00$		$\rho = 0.50$	
		Truth		
Y_1 alone				
AUC(0)	0.800	–	0.800	–
TP(FP = 0.20)	0.636	–	0.636	–
TP(FP = 0.10)	0.434	–	0.434	–
TP(FP = 0.05)	0.325	–	0.325	–
Optimal combination				
α_{opt}	1.00	–	1.00	–
AUC_{opt}	0.883	–	0.834	–
TP(FP = 0.20)	0.800	–	0.703	–
TP(FP = 0.10)	0.656	–	0.537	–
TP(FP = 0.05)	0.515	–	0.393	–
		$n^D = n^{\bar{D}} = 100$		
Distribution-free combination				
$\hat{\alpha}_{DF}$	1.00	(0.71, 1.41)	1.01	(0.53, 1.92)
$\widehat{AUC}_{DF}(\hat{\alpha}_{DF})$	0.887	(0.853, 0.915)	0.838	(0.799, 0.872)
AUC $(\hat{\alpha}_{DF})$	0.882	(0.877, 0.883)	0.833	(0.828, 0.834)
TP(FP = 0.20)	0.798	(0.789, 0.800)	0.701	(0.690, 0.703)
TP(FP = 0.10)	0.653	(0.641, 0.656)	0.534	(0.522, 0.537)
TP(FP = 0.05)	0.512	(0.500, 0.515)	0.391	(0.379, 0.393)
LDA combination				
$\hat{\alpha}_{LDA}$	1.00	(0.75, 1.36)	1.01	(0.57, 1.76)
$\widehat{AUC}_{LDA}(\hat{\alpha}_{LDA})$	0.885	(0.851, 0.912)	0.837	(0.797, 0.869)
AUC $(\hat{\alpha}_{LDA})$	0.882	(0.879, 0.883)	0.834	(0.829, 0.834)
TP(FP = 0.20)	0.798	(0.791, 0.800)	0.701	(0.693, 0.703)
TP(FP = 0.10)	0.654	(0.645, 0.656)	0.535	(0.526, 0.537)
TP(FP = 0.05)	0.513	(0.503, 0.515)	0.391	(0.383, 0.393)
		$n^D = n^{\bar{D}} = 50$		
Distribution-free combination				
$\hat{\alpha}_{DF}$	0.99	(0.59, 1.58)	0.970	(0.35, 2.42)
$\widehat{AUC}_{DF}(\hat{\alpha}_{DF})$	0.890	(0.843, 0.929)	0.844	(0.787, 0.891)
AUC $(\hat{\alpha}_{DF})$	0.881	(0.872, 0.883)	0.832	(0.821, 0.834)
TP(FP = 0.20)	0.797	(0.777, 0.800)	0.699	(0.676, 0.703)
TP(FP = 0.10)	0.651	(0.627, 0.656)	0.532	(0.507, 0.537)
TP(FP = 0.05)	0.510	(0.484, 0.515)	0.389	(0.364, 0.393)
LDA combination				
$\hat{\alpha}_{LDA}$	0.99	(0.64, 1.51)	0.968	(0.41, 2.26)
$\widehat{AUC}_{LDA}(\hat{\alpha}_{LDA})$	0.888	(0.839, 0.925)	0.841	(0.783, 0.885)
AUC $(\hat{\alpha}_{LDA})$	0.882	(0.874, 0.883)	0.833	(0.825, 0.834)
TP(FP = 0.20)	0.797	(0.782, 0.800)	0.699	(0.684, 0.703)
TP(FP = 0.10)	0.652	(0.633, 0.656)	0.533	(0.515, 0.537)
TP(FP = 0.05)	0.511	(0.490, 0.515)	0.390	(0.372, 0.393)

In essence we maximize a measure of distance between the distributions of the linear scores, $S(\alpha, Y) = \sum \alpha_p Y_p$, for diseased and non-diseased populations. Our method is similar to discriminant analysis in this regard. Discriminant analysis maximizes the ratio of between group variance to within group variance. We maximize the AUC or partial AUC which are different and general measures of distance between distributions that can be estimated using rank information only. Moreover, since ROC curves are well accepted as measures of accuracy in diagnostic medicine, it seems natural to use the area or partial area under the ROC curve as the objective function to be optimized. Su and Liu (1993) have previously argued this point and developed methodology for the case of multivariate binormal data. In this paper we have put forth a distribution-free methodology that is applicable more generally.

In addition to comparing our methodology with linear discriminant analysis, we have compared it with logistic regression, which also yields a linear score and which is widely available. In the various examples we considered, the LR procedure performed well. Nevertheless, since the estimated coefficients in the linear score are derived by maximizing a likelihood, the procedure does not appear to have a clear link to a relevant objective criterion for the diagnostic setting. Further investigation of logistic regression in this regard may be warranted. In addition there is no clear way of incorporating covariates that are only observed in the diseased subjects.

We focused here on settings where two markers are available, in part because both of our applications involved only two markers and in part because computation is relatively easy in this case. When $P > 2$ markers are involved, the problem is to find the $P-1$ coefficients $\alpha = \{\alpha_2, \ldots, \alpha_P\}$ such that the AUC or partial AUC for the score $S(\alpha, Y) = Y_1 + \sum_2^P \alpha_k Y_k$ is maximized. The DF approach can be applied as described in Section 2.3 but now a search in $(P-1)$-dimensional space for $\alpha = (\alpha_2, \ldots, \alpha_P)$ is required. This is straightforward but computationally demanding. As an alternative to searching simultaneously for $(\alpha_2, \ldots, \alpha_P)$, one might consider a stepwise approach. The first step would be to find the two markers whose optimal linear combination is best in the sense of having maximal AUC (or pAUC) amongst all pairs of markers. Having derived that score, and without loss of generality we denote it by $S^1(\alpha_2) = Y_1 + \alpha_2 Y_2$, the next step is to find the marker that when put in optimal linear combination with $S^1(\alpha_2)$ yields the best optimized AUC among all $P-2$ remaining markers. Without loss of generality we denote the optimized score by $S^2(\alpha_2, \alpha_3) = Y_1 + \alpha_2 Y_2 + \alpha_3 Y_3$. One can proceed in this fashion until all P markers are included in the linear combination. The advantage of the stepwise approach is that each step requires computation for only two markers at a time, and as described in Section 2.3 this requires the simple task of searching in two finite intervals in one-dimensional space, $\alpha \epsilon [-1, 1]$ and $1/\alpha \epsilon [-1, 1]$. The disadvantage is that the weights $(\alpha_2, \ldots, \alpha_P)$ derived in this fashion may not be optimal in $(P-2)$-dimensional space. Whether or not this is practically important remains to be explored.

The smooth distribution-free (SDF) approach can also be generalized to deal with $P > 2$ markers. Rather than one-dimension for α, there are $P-1$-dimensions, and thus for the (i, j) pair of diseased and non-diseased observations we create $(P-1)$-dimensional vectors $\{\alpha_1^{ij}, \ldots, \alpha_m^{ij}\}$ and correspondingly m binary random variables U_{ijk} indicating if the score $Y_1 + \sum_2^P \alpha_p Y_p$ calculated for the ith diseased observation is greater than that for the jth non-diseased observation. One models the AUC $(\alpha_k^{ij}) = E(U_{ijk})$ as logit $\big(E(U_{ijk})\big) = \beta(\alpha_k^{ij})$ as a smooth function in α_k^{ij}. Since α_k^{ij} is $P-1$ dimensional, multivariate regression splines will be necessary for modelling (Dierckx, 1993). After fitting the model, a search of the fitted model in $(P-1)$-dimensional space is needed to find the optimal $\{\alpha_2, \ldots, \alpha_P\}$. If these computations are overly complex for the user, a stepwise procedure can be applied to the SDF algorithm, in analogy with that described above for the DF algorithm, wherein one marker at a time is added to the linear combination.

Aside from computational issues, when the number of markers is large one needs to be concerned with potentially over-fitting the data. Neural networks, classification trees, etc. deal with this by incorporating

penalty functions. The sorts of penalty functions appropriate for AUC or pAUC optimization are, however, likely not the same as those for likelihood optimization. This remains an area for future development.

We have introduced a new approach to deriving linear combinations of markers that maximize the area or partial area under the ROC curve. Our procedures avoid modelling probability distributions of the data, in contrast to linear discriminant analysis, which is based on a binormal model, for the data. Yet, simulation studies suggest that the new procedures seem to be efficient with binormal data. This efficiency is reminiscent of the statistical efficiency of the Mann–Whitney U statistic for comparing two normal distributions (Hollander and Wolfe, 1973). An important advantage of the new procedures is that they can be used to maximize the partial AUC, whereas this is not necessarily accomplished by LDA unless the covariance matrices in the diseased and non-diseased populations are proportional (Su and Liu, 1993). Moreover, the new procedures accommodate covariates in a natural way and allow covariates to be specific to diseased observations. The new procedures are, however, computationally more demanding than is discriminant analysis or logistic regression. The DF method requires evaluation of $\widehat{\text{AUC}}$ at each α in the range of interest. The SDF method requires that the data be reconfigured as $m \times n^D \times n^{\bar{D}}$ records a potentially very large number which can adversely affect speed of computations.

Confidence intervals for the AUC(α) functions can be calculated using resampling methods. Pointwise intervals based on the non-parametric estimator are shown in Figure 4 for the pancreatic cancer data. These can be used informally to determine if the data support combining the markers linearly or if a single marker suffices. In Figure 4 for example, the optimal AUC, AUC($\alpha = 0.39$) is within the 90% confidence interval for AUC($\alpha = 0$). This suggests that Y_1 alone is sufficient. Formal testing of the hypothesis that AUC($\alpha = 0$) =AUC($\hat{\alpha}_{\text{opt}}$) would require estimating AUC($\hat{\alpha}_{\text{opt}}$) for each resampled dataset, which could be achieved with cross-validation methods although this would be highly demanding computationally.

Combining markers as linear scores was the focus of this paper. Linear combinations are intuitively best suited to discrimination if either very large values of any marker suggests presence of disease, or moderate increases in several markers suggest disease. Restricting attention to linear scores also made the problem more tractable. One could, however, consider a larger space of functions, such as generalized additive scores $S(\alpha) = g_1(Y_1) + \sum \alpha_p g_p(Y_p)$ where g_p is a linear combinations of (basis) functions of Y_p. The problem then is to optimize the AUC or partial AUC with respect to parameters in g_p, $p = 1, \ldots, P$ and with respect to $\alpha_2, \ldots, \alpha_P$. Although the maximization is essentially the same as that previously discussed for the multiple marker problem with $P > 2$, the multiplicity of parameters involved can make this particularly difficult computationally.

In summary, we have proposed methods for finding linear combinations of markers to maximize measures of accuracy which are commonly used in diagnostic medicine. The methods are distribution-free, appear to have good statistical properties, and can incorporate heterogeneity characterized by covariates. Further applications of the methods to real datasets will help define their potential role in practice.

ACKNOWLEDGEMENTS

This research was supported by grants R01 GM54438 and R01 HL57288 from the National Institutes of Health. We thank CARET study investigators for allowing us to use the PSA data and Sam Wieand for providing us with the prostate cancer data. We are most grateful to Molly Jackson and Gary Longton for assistance with manuscript preparation and to Ruth Etzioni for helpful comments on an earlier draft manuscript.

REFERENCES

BAMBER D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology* **12**, 387–415.

BEGG, C. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine* **10**, 1887–1895.

COLE, T. J. AND GREEN, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Statistics in Medicine* **11**, 1305–1319.

DAY, N. E. AND KERRIDGE, D. F. (1967). A general maximum likelihood discriminant. *Biometrics* **23**, 313–323.

DIERCKX, P. (1993). *Curve and Surface Fitting with Splines*. New York: Oxford University Press.

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

EFRON, B. (1975). The efficiency of logistic regression compared to normal discriminant analysis. *Journal of the American Statistical Association* **70**, 892–898.

EFRON, B. (1991). Regression percentiles using asymmetric squared error loss. *Statistica Sinica* **1**, 93–125.

ETZIONI, R., PEPE, M., LONGTON, G., HU C. AND GOODMAN, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: a prostate cancer case study. *Medical Decision Making* **19**, 242–251.

HANLEY, J. A. AND MCNEIL, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology* **143**, 29–36.

HE, X. (1997). Quantile curves without crossing. *The American Statistician* **51**, 186–192.

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

HOLLANDER, M. AND WOLFE, D. A. (1973). *Nonparametric Statistical Methods*. New York: John Wiley & Sons.

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

MCCLISH, D. K. (1989). Analyzing a portion of the ROC curve. *Medical Decision Making* **9**, 190–195.

OBUCHOWSKI, N. A. (1995). Multireader, multimodality receiver operating characteristic curve studies: hypothesis testing and sample size estimation using an analysis of variance approach with dependent observations. *Academic Radiology* **2** (suppl. 1), S22–S29.

PEARSON, J. D., LUDERER, A. A., METTER, E. J., PARTIN, A. W., CHAN, D. W., FOZARD, J. L. AND CARTER, H. B. (1996). Longitudinal analysis of serial measurements of free and total PSA among men with and without prostatic cancer. *Urology* **48(6A)**, 4–9.

PEPE, M. S. (1997). A regression modelling framework for ROC curves in medical diagnostic testing. *Biometrika* **84**, 595–608.

PEPE, M. S. (1998). Three approaches to regression analysis of receiver operating characteristic curves for continuous test results. *Biometrics* **54**, 444–452.

RICHARDS, R. J., HAMMITT, J. K. AND TSEVAT, J. (1996). Finding the optimal multiple-test strategy using a method analagous to logistic regression. *Medical Decision Making* **16**, 367–375.

SU, J. Q. AND LIU, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350–1355.

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.

THORNQUIST, M. D., *et al*. (1993). Statistical design and monitoring of the Carotene and Retinol Efficacy Trial (CARET). *Controlled Clinical Trials* **14**, 308–324.

TOSTESON, A. N. A. AND BEGG, C. B. (1988). A general regression methodology for ROC curve estimation. *Medical Decision Making* **8**, 204–215.

WIEAND, S., GAIL M. H., JAMES, B. R. AND JAMES, K. L. (1989). A family of nonparametric statistics for comparing diagnostic markers with paired or unpaired data. *Biometrika* **76**, 585–592.

A. APPENDIX

A.1. Implementing the smooth distribution-free (SDF) approach

A key step in implementing the SDF method concerns the choice of relative weights $\{\alpha_1^{ij}, \ldots, \alpha_m^{ij}\}$. We chose half of the weights $\{\alpha_1^{ij}, \ldots, \alpha_{m/2}^{ij}\}$ in the interval $(-1, 1)$ and the remainder $\{\alpha_{m/2+1}^{ij}, \ldots, \alpha_m^{ij}\}$ such that $\gamma_k^{ij} = (\alpha_k^{ij})^{-1}$ are located in $(-1, 1)$. As described for the DF approach, this makes the procedure symmetric in Y_1 and Y_2. Since the objective is to fit the AUC(α) model over the entire domain $\alpha\epsilon(-\infty, \infty)$ it is important that the design points, i.e. the αs, cover the domain. In our illustrations we chose $\{\alpha_1^{ij}, \ldots, \alpha_{m/2}^{ij}, \gamma_{m/2+1}^{ij}, \ldots, \gamma_m^{ij}\}$ to have independent uniform distributions in $(-1, 1)$. Even if m, the number of α's chosen for each ij pair, is small, if there are a large number of such pairs, the overall effect is to densely cover the domain.

To fit a smooth function to AUC(α) in $(-1, 1)$, we parameterized it as a cubic regression spline with fixed knots in $(-1, 1)$. In order to deal with $\alpha \notin (-1, 1)$, one could similarly define for $\alpha^{-1} = \gamma \in (-1, 1)$ a regression spline for AUC(γ^{-1}). However, to ensure continuity of the AUC at $\alpha = \gamma = 1$ we modified the procedure as follows. We considered the range $(-1, 3)$, where $(-1, 1)$ was relevant to $\alpha \in (-1, 1)$ and the range $(1, 3)$ was relevant to $2 - \gamma$ where $\gamma \in (-1, 1)$. Choose a set of P knots in the range $(-1, 3)$ and calculate the corresponding basis functions for α_k^{ij} if $\alpha_k^{ij} \in (-1, 1)$ and for $(2 - \gamma_k^{ij})$ if $\gamma_k^{ij} = (\alpha_k^{ij})^{-1} \in (-1, 1)$. The logistic regression model for the binary variables U_{ijk} is fit to these basis functions of α_k^{ij}, resulting in a smooth functional form for AUC(α) which is continuous at $\alpha = 1$. In our examples we choose $P = 17$ knots for the regression spline model located at $\{-0.9, -0.75, -0.50, -0.25, 0, 0.25, 0.50, 0.75, 1.00, 1.25, 1.50, 1.75, 2.00, 2.25, 2.50, 2.75, 2.9\}$.

A.2. Implementing the partial AUC maximization

The key aspect not described in the main text for applying the SDF approach to maximizing the partial AUC, was estimation of the $1 - t_0$ quantiles associated with $Y_1^{\bar{D}} + \alpha\, Y_2^{\bar{D}}$, denoted by $Q^{\bar{D}}(1 - t_0, \alpha)$. We choose to implement a semiparametric method due to Heagerty and Pepe (1999), although alternative quantile regression methods such as those of Koenker and Bassett (1978), He (1997), Cole and Green (1992) or Efron (1991) might be used instead. Heagerty and Pepe (1999) assume that the data follow a location and scale family with unspecified baseline distribution function. We construct $m \times n^D \times n^{\bar{D}}$ data records, $\{S_{ijk} = Y_{j1} + \alpha_k^{ij} Y_{j2}$ for $\alpha_k^{ij} \in (-1, 1)$ and $S_{ijk} = (\alpha_k^{ij})^{-1}\, Y_{j1} + Y_{j2}$ for $\alpha_k^{ij} \notin (-1, 1); k = 1, \ldots, m;\ i = 1, \ldots, n^D;\ j = 1, \ldots, n^{\bar{D}}\}$ and model the mean and variance as smooth functions of α. Specifically, as described in A.1 we use a regression spline on $(-1, 3)$ for the mean function μ and for the variance function σ^2, with the range $(-1, 1)$ pertaining to $\alpha_k^{ij} \in (-1, 1)$ and the range $(1, 3)$ pertaining to $(2 - (\alpha_k^{ij})^{-1})$ for $\alpha_k^{ij} \notin (-1, 1)$. The empirical quantiles of the residuals, $\{(S_{ijk} - \hat{\mu}(\alpha_k^{ij}))/\hat{\sigma}(\alpha_k^{ij}); i = 1, \ldots, n^D;\ j = 1, \ldots, n^{\bar{D}};\ k = 1, \ldots, m\}$, denoted by $\hat{Q}_0(q)$ for $q \in (0, 1)$, are used to estimate quantiles of the baseline distribution. The estimated quantile $\hat{Q}^{\bar{D}}(1 - t_0, \alpha_k^{ij})$ is then given by $\hat{\mu}(\alpha_k^{ij}) + \hat{\sigma}(\alpha_k^{ij})\hat{Q}_0(1 - t_0)$.

[*Received April 26, 1999*; *revised November 1, 1999*; *accepted for publication November 22, 1999*]