

## 1. INTRODUCTION

The use of clinical and laboratory data to detect conditions and predict patient outcomes is a mainstay of medical practice. Classification and prediction are equally important in other fields of course (e.g., meteorology, economics, computer science) and have been subjects of statistical research for a long time. The field is currently receiving more attention in medicine, in part because of biotechnologic advancements that promise accurate non-invasive modes of testing. Technologies include gene expression arrays, protein mass spectrometry and new imaging modalities. These can be used for purposes such as detecting subclinical disease, evaluating the prognosis of patients with disease and predicting their responses to different choices of therapy. Statistical methods have been developed for assessing the accuracy of classifiers in medicine (Zhou, Obuchowski, and McClish 2002; Pepe 2003), although this is an area of statistics that is evolving rapidly.

In practice there may be multiple sources of information available to assist in prediction. For example, clinical signs and symptoms of disease may be supplemented by results of laboratory tests. As another example, it is expected that multiple biomarkers will be needed for detecting subclinical cancer with adequate sensitivity and specificity (Pepe et al. 2001). The starting point for the research we describe in this paper is the need to combine multiple predictors together, somehow, in order to predict outcome. We note in Section 2 that for a binary outcome,  $D$ , the best classifiers using the predictors  $Y = (Y_1, Y_2, \dots, Y_P)$  are based on the risk score function

$$RS(Y) = P(D = 1|Y) \tag{1}$$

or equivalently any monotone increasing transformation of it. Rules based on  $RS(Y) > c$  are optimal, where the choice of threshold  $c$  is determined by the stringency of the criterion that is acceptable in the setting where classification is to take place, a topic not addressed in this paper.

Suppose that we employ a generalized linear model for the risk score

$$P(D = 1|Y) = g(\alpha_0 + \alpha_1 Y_1 + \dots + \alpha_P Y_P) \tag{2}$$

where  $g^{-1}$  is a monotone increasing link function and without loss of generality we assume that the predictors are coded so that the coefficients are positive. The fixed set of  $P$  predictors can be as many and as flexible as desired, including interactions and mathematical functions of the predictive variables available. The specific statistical question we address in this paper concerns the estimation of the linear predictor, once  $(Y_1, \dots, Y_P)$  have been chosen.

The logistic likelihood can be maximized in order to estimate the coefficients  $(\alpha_1, \dots, \alpha_P)$ . In Section 2 we propose an alternative objective function, the AUC, that does not require specification of the link function  $g$ , that applies under a variety of sampling schemes and that provides useful results even when the generalized linear model does not hold. Simulation studies presented in Section 3 explore its performance relative to the logistic likelihood. Although the AUC performs well under a variety of circumstances, contrary to our expectations, the logistic likelihood appears to be similarly robust. Possible explanations for this, which derive from recent work by Eguchi and Copas (2002), are outlined in Section 4.

In the latter parts of the paper we propose that in practice the risk score need only be estimated over a subset of the predictor space. This can greatly reduce the scope of modeling assumptions made, thereby inducing further robustness to the analysis. The logistic likelihood objective function is compared with the restricted region analogue of the AUC (the partial AUC) for model fitting. Again their performances in simulation studies are found to be similar across a range of settings. We close in Section 7 with a discussion of the practical implications of our results and recommendations for further statistical research in this area.

## 2. BINARY REGRESSION

## 2.1 Optimality of the Risk Score Function

Consider a rule that classifies subjects as positive for the class ( $D = 1$ ) or negative. The operating characteristics of such a rule are its true and false positive fractions, TPF and FPF, respectively, where

$$\text{TPF} = P[\text{positive} | D = 1]$$

and

$$\text{FPF} = P[\text{positive} | D = 0].$$

These are also known as the sensitivity (TPF) and specificity ( $1 - \text{FPF}$ ) of the rule. Among all rules with a fixed FPF, the best rule is that with maximum TPF. Similarly with TPF fixed, the best rule is that with minimum FPF. It can be shown using the Neyman-Pearson lemma (Neyman and Pearson 1933) that the optimal rules in this sense have the form

$$'RS(Y) > c,' \tag{3}$$

where  $c$  is chosen to fix the FPF or TPF as required. See Green and Swets (1966) or, for a more recent exposition, see McIntosh and Pepe (2002). It can be shown that, as a direct corollary, these rules also minimize the misclassification rate and they minimize the expected cost of false positive and false negative errors combined. Bayesians have long promoted the risk score function because of these latter two properties.

Pictorially, the Neyman-Pearson result states that the risk score has the optimal receiver operating characteristic (ROC) curve. The ROC curve plots the TPF of the rule ' $RS(Y) > c$ ' versus its FPF for all possible choices of threshold  $c$  (Figure 1). Neyman-Pearson implies that any rule based on the predictors  $Y = (Y_1, \dots, Y_P)$  cannot have a classification probability point (FPF, TPF) lying above the ROC curve for the risk score function. Recently Eguchi and Copas (2002) and Baker (2000) noted this optimality property for the likelihood ratio function

$\mathcal{LR}(Y) = P(Y|D = 1)/P(Y|D = 0)$ . Since the risk score,  $\text{RS}(Y)$ , is a monotone increasing function of  $\mathcal{LR}(Y)$ , rules based on their exceeding thresholds are equivalent. The ROC curve for  $\text{RS}(Y)$  is therefore the same as that of  $\mathcal{LR}(Y)$ . We use the risk score formulation here because approaches for estimating it are more familiar than are procedures for estimating the likelihood ratio function.

## 2.2 The Logistic Likelihood Approach

Suppose that we have data for  $n_D$  observations truly classified as  $D = 1$  and for  $n_{\bar{D}}$  observations with true class  $D = 0$ . We write the data as  $\{Y_{D1}, \dots, Y_{Dn_D}\}$  and  $\{Y_{\bar{D}1}, \dots, Y_{\bar{D}n_{\bar{D}}}\}$ . Frequently, sampling will depend on the true classification status. We call this retrospective sampling. This is particularly true for studies of relatively rare conditions or outcomes in medicine where case-control designs are very common (Pepe et al. 2001). Logistic regression is popular for case-control designs because the regression parameters  $(\alpha_1, \dots, \alpha_P)$  in (2) can be estimated consistently from the simple prospective log-likelihood,

$$\log \mathcal{L}(\alpha) = \sum_{i=1}^{n_D} \log P(D_i = 1|Y_{Di}) + \sum_{j=1}^{n_{\bar{D}}} \log P(D_j = 0|Y_{\bar{D}j})$$

even when sampling is retrospective (Prentice and Pyke, 1979). With logistic link function  $g^{-1}(X) = \log\{X/(1 - X)\}$ , the log-likelihood is

$$\log \mathcal{L}^L(\alpha) = \sum_{i=1}^{n_D} \alpha Y_{Di} - \sum_{k=1}^{n_D+n_{\bar{D}}} \log(1 + e^{\alpha Y_k}) \quad (4)$$

where  $\alpha Y = \alpha_0 + \alpha_1 Y_1 + \dots + \alpha_P Y_P$  and the second summation is over all  $n_D + n_{\bar{D}}$  observations.

We denote the logistic maximum likelihood estimates as  $\hat{\alpha}^L = (\hat{\alpha}_0^L, \hat{\alpha}_1^L, \dots, \hat{\alpha}_P^L)$ .

The corresponding estimated risk score function is  $\widehat{\text{RS}}^L(Y) = g(\hat{\alpha}^L Y)$ . When a case-control design is employed, however, the estimated intercept,  $\hat{\alpha}_0^L$  is not consistent although because the logistic link is employed  $(\hat{\alpha}_1^L, \dots, \hat{\alpha}_P^L)$  are consistent, (Prentice and Pyke 1979). Therefore  $\widehat{\text{RS}}^L(Y)$  is not consistent for the true risk score. Fortunately, it is not necessary to estimate the risk score

itself. Any monotone increasing transformation of  $RS(Y)$  has the same ROC curve as  $RS(Y)$ . In particular, the linear predictor

$$\alpha_1 Y_1 + \dots + \alpha_P Y_P = g^{-1}(RS(Y)) - \alpha_0$$

is such a monotone increasing transformation and it is therefore enough to estimate it. We go one step further and note that the linear predictor rescaled by any constant is also equivalent to the optimal function  $RS(Y)$ . For identifiability we set the first coefficient to 1 by dividing by  $1/\alpha_1$  and define the rescaled linear predictor

$$l_\beta(Y) = Y_1 + \beta_2 Y_2 + \dots + \beta_P Y_P \tag{5}$$

where  $\beta_p = \alpha_p/\alpha_1$ . In summary, these arguments indicate that under the logistic model this linear predictor optimally combines  $(Y_1, \dots, Y_P)$  for the purpose of classifying future observations as  $D = 1$  or  $D = 0$ . Using the data on  $n_D$  cases and  $n_{\bar{D}}$  controls, the logistic regression procedure provides the estimated linear predictor function that will be used for classification

$$\widehat{l}_\beta^L(Y) = Y_1 + \widehat{\beta}_2^L Y_2 + \dots + \widehat{\beta}_P^L Y_P$$

where  $\widehat{\beta}_p^L = \widehat{\alpha}_p^L/\widehat{\alpha}_1^L$ .

### 2.3 The AUC Approach

The arguments of the previous section about the need to estimate only the linear predictor,  $l_\beta(Y)$  as opposed to  $RS(Y)$  do not rely on the link function  $g$  having logistic form. They apply to the general model (2). Another approach to estimating  $\beta = (\beta_2, \dots, \beta_P)$  is motivated as follows. Since  $l_\beta(Y)$  has the best ROC curve among all functions of  $Y$ , it certainly has the best ROC curve among all linear predictors of the form

$$l_b(Y) = Y_1 + b_2 Y_2 + \dots + b_P Y_P.$$

The idea is to select choices of coefficients  $(b_2, \dots, b_P)$  that yield the best observed ROC curve. These are then interpreted as estimates of  $(\beta_2, \dots, \beta_P)$ .

The area under the ROC curve (AUC) is the most popular ROC summary index. The optimal ROC curve has the maximum AUC, so we can use it as the basis for an objective function of the data to estimate  $\beta$ . The empirical estimator of the AUC is the Mann-Whitney U statistic

$$\widehat{\text{AUC}}(b) = \frac{\sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} I[l_b(Y_{Di}) > l_b(Y_{\bar{D}j})]}{n_D n_{\bar{D}}}.$$

We write the AUC based estimator of  $\beta$  as

$$\widehat{\beta}^{\text{AUC}} = \operatorname{argmax}(\widehat{\text{AUC}}(b)).$$

This can be recognized as a special case of the maximum rank correlation estimator described by Han (1987). The estimator is known to be consistent and asymptotically normal under the generalized linear model (2). See Sherman (1993) for these results.

#### 2.4 Relative merits theoretically

One major attribute of the AUC approach is that it does not require the link function,  $g$ , to be specified. It works regardless of the form of the true link function  $g$ . On the other hand, the logistic approach depends on the assumption that  $g$  is logistic,  $g(X) = e^X/(1 + e^X)$ . Presumably the logistic regression procedure might fail when  $g$  is not logistic.

The logistic model is popular over other forms for  $g$ , in part because it accommodates either prospective or retrospective (case-control) designs. Interestingly, we see that the AUC approach shares this property. Because the AUC conditions on the binary response variables  $\{D_i = 1, i = 1, \dots, n_D; D_j = 0, j = 1, \dots, n_{\bar{D}}\}$ , it allows sampling to depend on  $D$ . Thus, it accommodates case-control designs too. Moreover, unlike logistic regression, it does so without restricting the form of the link function to be logistic.

Finally, and most importantly, we consider the two approaches when the generalized linear

model (2) does not hold. The AUC approach still yields a sensible entity, namely the linear combination,  $l_b(Y)$ , that maximizes the area under the ROC curve. Even though the resulting linear predictor may not have the optimal ROC curve associated with  $RS(Y)$ , it has the optimal one among all linear combinations of the predictors. In contrast, there are no obvious optimality properties for the linear predictor derived from the logistic likelihood when (2) fails in general.

The one theoretical advantage of the logistic approach over the AUC approach is its asymptotic efficiency when (2) holds and  $g$  is logistic. In the next section we assess the relative efficiency of the methods in this setting. Then we investigate if the flexibility of the AUC approach offered by the properties of not requiring specification of a link function and having validity when (2) fails, translate into practically meaningful benefits.

### 3. SIMULATION STUDIES

We simulated a variety of models and report results for a representative subset here. Predictor variables  $(Y_1, Y_2)$  were generated with asymmetric and non-independent distributions. Specifically, we generated independent random variables  $(Y, \tilde{Y}_1, \tilde{Y}_2)$  with distributions  $Y \sim \text{Gamma}(3,5)$ ,  $\tilde{Y}_1 \sim N(10,4)$  and  $\tilde{Y}_2/10 \sim \text{Gamma}(2,3)$  and calculated

$$Y_1 = 0.5Y + 0.5\tilde{Y}_1 \quad Y_2 = 0.5Y + 0.5\tilde{Y}_2.$$

#### 3.1 Linear logistic risk score

To assess the relative efficiency of the maximum logistic likelihood estimator compared to the maximum AUC estimator, we first simulated the binary response,  $D$ , with a logistic model:

$$\text{logit}P(D = 1|Y_1, Y_2) = -6 + Y_1 + \beta_2 Y_2.$$

with  $\beta_2 = 0.5$ . Then we used a case-control sampling scheme to select  $n_D$  cases and  $n_{\bar{D}}$  controls. Estimation of  $\beta_2$  from both methods are compared in Table 1.

The top two rows show that although  $\beta_2$  is estimated with little bias by both methods, the maximum likelihood method provides substantially more efficient estimates of  $\beta_2$ . In terms of the classification probabilities associated with the estimated linear predictor, that derived from the logistic likelihood appears to be somewhat better also. The area under its ROC curve ( $AUC(\hat{\beta}_2^L)$ ) is on average the same as that for  $AUC(\hat{\beta}_2^{AUC})$  but it is somewhat less variable. For assessing performance of the linear predictor it is probably more relevant to calculate a measure such as the root-mean squared distance between the AUC of the linear predictor from the optimal AUC value of 0.808 associated with the ‘true’ linear predictor,  $Y_1 + \beta_2 Y_2 = Y_1 + 0.5Y_2$ . Results in the last two rows of Table 1 can be used to calculate this. For example, with  $n_D = 100$  and  $n_{\bar{D}} = 1000$  we have

$$\begin{aligned}\sqrt{E[AUC(\hat{\beta}_2^L) - AUC(0.5)]^2} &= \sqrt{(.006)^2 + (.0014)^2} = .0062 \\ \sqrt{E[AUC(\hat{\beta}_2^{AUC}) - AUC(0.5)]^2} &= \sqrt{(.006)^2 + (.0021)^2} = .0064\end{aligned}$$

On this scale the difference between the two methods is minor. This holds for other sample sizes and data configurations also. We conclude that when the linear logistic model holds the performance of the linear predictor derived from maximizing the Mann-Whitney U statistic is comparable with the statistically efficient maximum likelihood derived linear predictor.

### 3.2 Misspecified link function

We next compared performances when the link function  $g$  is not logistic. We simulated data from the model

$$P(D = 1|Y_1, Y_2) = L\{H(-6 + Y_1 + 0.5Y_2)\}$$

with  $L(\cdot)$  corresponding to the cumulative distribution function of a mixture of a Weibull (3,3) distribution with probability 0.5 and Weibull (4,11) with probability 0.5. Two different forms for  $H$

were used:  $H_1(x) = \exp(x)$  and  $H_2(x) = \exp\{(x+8)/3\}$ . To visualize how far away  $g(\cdot) = L(H(\cdot))$  is from the logistic link, we examine the linearity of the function  $\text{logit}\{L(H(x))\}$  in  $x$  for the relevant range of  $x$ , corresponding to the linear predictors in our simulation. As shown in Figure 2, the misspecification is fairly severe. Results shown in Table 2 for the first setting yield conclusions similar to those found earlier in Table 1 when the logistic link holds. In the second setting where  $H(x) = \exp((x+8)/3)$ , we find that  $\widehat{\beta}_2^L$  can be substantially biased and interestingly the bias varies with the ratio of cases to controls,  $n_D/n_{\bar{D}}$ , in the sample. In contrast, the robust estimator  $\widehat{\beta}_2^{\text{AUC}}$  that does not require specification of the link function is relatively unbiased compared with  $\widehat{\beta}_2^L$ . However the bias in  $\widehat{\beta}_2^L$  does not translate into meaningful reductions in the performance of the linear predictor  $Y_1 + \widehat{\beta}_2^L Y_2$  relative to  $Y_1 + \widehat{\beta}_2^{\text{AUC}} Y_2$ .

### 3.3 Misspecified linear predictor

To further explore the robustness of the AUC approach compared to the logistic likelihood approach we simulated data from the following model

$$\text{logit } P(D = 1|Y_1, Y_2) = -7 + Y_1 + \left(\frac{1}{3}Y_1 - 1\right)Y_2. \quad (6)$$

We assumed the linear form  $Y_1 + \beta_2 Y_2$  for the predictor and calculated the estimates  $\widehat{\beta}_2^L$  and  $\widehat{\beta}_2^{\text{AUC}}$  accordingly. Thus the linear predictor assumed in the estimation is a misspecification in this setting. Although the estimators are not necessarily converging to the same limits (see the mean values when  $n_D = 1000$  and  $n_{\bar{D}} = 100$  in Table 3), the associated linear predictors appear to perform equally well. Moreover, we note that the optimal combination,  $Y_1 + (\frac{1}{3}Y_1 - 1)Y_2$ , has an AUC of 0.895, so that the linear combinations perform well but not optimally in this setting.

In another study, instead of simulating data based on generalized linear models, we simulated  $(Y_1, Y_2)$  from a mixture of normal distributions as described by McIntosh and Pepe (2002). Specifically, we simulate  $Y_1$  and  $Y_2$  from independent standard normal distribution if  $D = 0$  and if  $D = 1$ , we simulate from a mixture of four bivariate normal distributions:

$$\begin{aligned}
& N \left( \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1.00 & 0 \\ 0 & 1.00 \end{bmatrix} \right) \text{ with prob}=0.012, & N \left( \begin{bmatrix} 0 \\ 2.54 \end{bmatrix}, \begin{bmatrix} 1.00 & 0 \\ 0 & 0.25 \end{bmatrix} \right) \text{ with prob} = 0.188, \\
& N \left( \begin{bmatrix} 1.68 \\ 0 \end{bmatrix}, \begin{bmatrix} 0.04 & 0 \\ 0 & 1.00 \end{bmatrix} \right) \text{ with prob}=0.689, & N \left( \begin{bmatrix} 1.68 \\ 2.54 \end{bmatrix}, \begin{bmatrix} 0.04 & 0.09 \\ 0.09 & 0.25 \end{bmatrix} \right) \text{ with prob} = 0.111.
\end{aligned}$$

This is another setting where the risk score is not well approximated by a monotone transformation of a linear predictor of the form  $Y_1 + \beta_2 Y_2$ . Nevertheless we again find from Table 4 that the performance of the linear predictor derived from the logistic likelihood compares well with that derived from the optimal AUC statistic even though the logistic linear model does not hold.

#### 4. ROBUSTNESS OF THE LOGISTIC APPROACH

The simulation results are both encouraging and puzzling. They are encouraging because they indicate the popular widely available statistical procedure, that is logistic regression, produces linear predictors that are robust.

Li and Duan (1989) have previously shown that logistic regression is robust to misspecification of the true link function  $g$  as a logistic form. They show that under some technical conditions on the distributions of predictors, the rescaled coefficients  $(\beta_2, \dots, \beta_P)$  are estimated consistently by logistic regression when the true link function is not logistic. We had considerable difficulty identifying a setting where the Li and Duan condition failed, but appear to have done so in the second simulation study of section 3.2. The estimator  $\hat{\beta}_2^L$  is biased in large samples (lower panel of Table 2). Nevertheless the linear predictor,  $Y_1 + \hat{\beta}_2^L Y_2$ , performs as well as the predictor based on  $\hat{\beta}_2^{\text{AUC}}$  which is consistent for  $\beta_2$ . This is puzzling.

Even more puzzling to these authors is the fact that the AUC associated with  $Y_1 + \hat{\beta}_2^L Y_2$  is as good as that based on  $Y_1 + \hat{\beta}_2^{\text{AUC}} Y_2$  when the linear form of the risk score is misspecified. Why

does maximizing the logistic likelihood objective function yield a linear combination with operating characteristics that are similar to those of the linear combination that maximizes the area under the ROC curve? Some clues may be found in a recent paper by Eguchi and Copas (2002). They show that objective functions that yield consistent results under the logistic model can be approximated by the weighted integral of the difference between the ROC curve associated with the true risk score (or likelihood ratio) and that associated with the linear combination. Since the logistic likelihood is certainly an objective function in this class, we therefore can approximate it as

$$\log \mathcal{L}^L(b) = \int_0^1 w(t) \{ \text{ROC}(b, t) - \text{ROC}^{\text{opt}}(t) \} dt$$

where  $t$  is the false-positive fraction index,  $\text{ROC}(b, t)$  is the associated TPF for the linear combination  $Y_1 + bY_2$  and  $\text{ROC}^{\text{opt}}(t)$  is the associated TPF for the risk score. If the weight  $w(t)$  were constant in  $t$  then maximizing  $\mathcal{L}^L(b)$  therefore amounts to maximizing  $\int_0^1 \text{ROC}(b, t) dt = \text{AUC}(b)$ . This result may hold approximately when  $w(t)$  is not constant.

## 5. RESTRICTING THE PREDICTOR SPACE

### 5.1 Rationale

Classification rules must achieve operating characteristics that are acceptable for the settings in which they are applied. In disease screening of healthy populations large false positive fractions are unacceptable since they lead to large numbers of healthy people being unnecessarily referred for invasive work-up or treatment. Moreover, screening tests must detect a reasonable fraction of people with disease in order to be worthwhile. Similar considerations are relevant for medical tests applied in other settings. Let  $f_0$  denote the largest acceptable false positive fraction for the applications envisioned, and let  $t_0$  denote the smallest acceptable true positive fraction. That is to say, that classification rules with false positive fractions larger than  $f_0$  or true positive fractions smaller than  $t_0$  are not of interest. Pictorially, only the region of ROC space shaded in Figure 3(A)

is of concern.

The optimal rules in this region require estimating the risk score function only over the predictor subspace that corresponds to this region. This obviously reduces the complexity of the task over estimating  $RS(Y)$  over the whole predictor space. Observe in Figure 2(B) that the restricted ROC region corresponds to the predictor space region with values of  $RS(Y)$  between  $q_{\bar{D}}(f_0)$  and  $q_D(t_0)$  where the quantiles  $q_{\bar{D}}(f_0)$  and  $q_D(t_0)$  are the  $(1 - f_0)$  quantile of  $RS(Y)$  in the population  $D = 0$  and the  $(1 - t_0)$  quantile in the population  $D = 1$ , respectively. If  $q_{\bar{D}}(f_0)$ ,  $q_D(t_0)$  and  $RS(Y)$  are known then consider the rule that classifies

$$\text{Positive if } RS(Y) > q_D(t_0)$$

$$\text{Negative if } RS(Y) < q_{\bar{D}}(f_0)$$

and otherwise,

$$\text{Positive if } RS(Y) > q_{\bar{D}}(f)$$

where  $q_{\bar{D}}(f)$  is the  $1 - f$  quantile of  $RS(Y)$  in the population with  $D = 0$ . This rule has (FPF, TPF) values in the acceptable range and is optimal among all rules with  $FPF = f$  because for each  $f$  it is the Neyman-Pearson rule. This motivates estimating the  $q_D(t_0)$  and  $q_{\bar{D}}(f_0)$  quantiles of the risk score and the risk score function itself over the subspace of the predictor space  $\{Y : RS(Y) \in (q_{\bar{D}}(f_0), q_D(t_0))\}$ . Recall that under the generalized linear model (2), instead of the risk score function, we need only be concerned with the linear predictor  $l_\beta(Y)$ . The quantiles of the linear predictor, denoted by  $(\Psi_{\bar{D}}(f_0), \Psi_D(t_0))$  and the linear predictor function itself over the subspace, yield the same classification rules as  $(q_{\bar{D}}(f_0), q_D(t_0))$  and  $RS(Y)$ . A crude variation of this was proposed by McIntosh and Pepe(2002), using the logistic likelihood for estimation purposes. Here we employ a technique related to the AUC as well.

## 5.2 Two-step estimation

McIntosh and Pepe (2002) used the following sort of two-step procedure. First a linear logistic model for  $RS(Y)$  is fit over the whole predictor space using all  $n_D + n_{\bar{D}}$  observations. The estimated linear predictors are calculated, denoted by  $l_{\hat{\beta}_1}(Y)$  and the empirical quantiles  $(\hat{\Psi}_D^1(t_0), \hat{\Psi}_D^1(f_0))$  are calculated using the empirical distributions of  $\{l_{\hat{\beta}_1}(Y_{iD}), i = 1, \dots, n_D\}$  and  $\{l_{\hat{\beta}_1}(Y_{j\bar{D}}), j = 1, \dots, n_{\bar{D}}\}$ , respectively. In the second step, only observations  $(D_k, Y_k)$  that satisfy  $\hat{\Psi}_D^1(f_0) < l_{\hat{\beta}_1}(Y_k) < \hat{\Psi}_D^1(t_0)$  are included, and the linear logistic model is fit again. One could iterate the procedure, refining the estimates of  $(\Psi_{\bar{D}}(f_0), \Psi_D(t_0))$  thereby adjusting the predictor subspace and then refitting the model over it.

Instead of logistic regression, the AUC procedure can be used to fit the linear predictors at each step. When restricted to the predictor subspace,  $\{Y : \Psi_{\bar{D}}(f_0) \leq l_{\beta}(Y) \leq \Psi_D(t_0)\}$ , the AUC is more properly referred to as the rescaled area under the partial ROC curve (rpAUC). We write

$$\text{rpAUC} = \int_{\text{ROC}^{-1}(t_0)}^{f_0} (\text{ROC}(f) - t_0) df / (f_0 - \text{ROC}^{-1}(t_0))(1 - t_0)$$

and estimate it empirically with the Mann-Whitney statistic

$$\text{rp}\widehat{\text{AUC}}(b) = \sum_{i=1}^{n_D^R} \sum_{j=1}^{n_{\bar{D}}^R} \frac{I[l_B(Y_{Di}) \geq l_b(Y_{\bar{D}j})]}{n_D^R n_{\bar{D}}^R}$$

where  $n_D^R$  and  $n_{\bar{D}}^R$  denote the numbers of observations from each of the two samples,  $D = 1$  and  $D = 0$ , that lie in the restricted predictor space. We write  $\hat{\beta}^{\text{rpAUC}} = \text{argmax}(\text{rp}\widehat{\text{AUC}}(b))$  and  $\hat{\beta}^{rL}$  for the corresponding two-step estimator derived from the logistic likelihood.

Note that the predictors are coded so that the coefficients  $\beta$  are positive. Therefore the boundaries of the restricted space defined by  $\Psi_{\bar{D}}(f_0)$  and  $\Psi_D(t_0)$  define a contiguous region and smaller values of any predictor are associated with smaller values of the linear predictor (and of the risk score) in the restricted region. If we can assume that the risk score is smooth and monotone increasing in each predictor over the whole predictor space then the linear predictor function, which strictly speaking applies to the restricted region only, nevertheless orders the whole space ade-

quately for the purposes of estimating the quantiles  $(\Psi_{\bar{D}}(f_0), \Psi_D(t_0))$  and hence the boundaries of the restricted region. We implicitly make the monotonicity assumption here and suspect that in most practical applications it will hold. Modifications of the procedure to accommodate more complex risk score functions could be developed.

## 6. MORE SIMULATIONS

To compare the restricted region two-step procedures to the full region analyses when model (2) does not hold over the whole space, we again used the normal mixture setting to simulate data. We considered the ROC region of interest to be that where  $\text{FPF} \leq 0.3$ . The resulting estimates are summarized in Table 5. We find that the approach of maximizing the restricted region logistic likelihood and of maximizing  $\widehat{\text{rpAUC}}$  result in slightly different estimates of  $\widehat{\beta}_2$ . It is interesting to see that in this setting maximizing the rpAUC seems to produce a slightly better linear combination than that produced by the logistic likelihood over the restricted region ( $\text{rpAUC}(\widehat{\beta}_2^{rL}) = 0.786$  versus  $\text{rpAUC}(\widehat{\beta}_2^{\text{rpAUC}}) = 0.812$  when  $n_D = n_{\bar{D}} = 100$ ). More importantly these two approaches are both superior to the procedures that fit the linear model over the full prediction space, in the sense that their operating characteristics over the relevant region of ROC space are better. For example,  $\text{rpAUC}(\widehat{\beta}_2^L) = 0.748$  versus  $\text{rpAUC}(\widehat{\beta}_2^{rL}) = 0.786$  at  $n_D = n_{\bar{D}} = 100$ . That is, the restricted region procedures produce larger values of the ROC summary index rpAUC, for the linear predictors. This affirms the motivating rationale provided for estimating the linear predictor over the restricted region of the predictor space.

## 7. DISCUSSION

In this paper we have proposed a robust approach to deriving linear combinations of predictors for a binary outcome. Although our simulation studies indicate that logistic regression works equally well, there are some reasons to advocate in favor of using the estimated AUC (i.e., Mann-Whitney U

statistic) as the objective function. The main one is that there is rigorous theory to show that  $\widehat{\beta}^{\text{AUC}}$  is consistent and asymptotically normal under the generalized linear model (2). Although Li and Duan (1989) have shown consistency for logistic regression too, their results require conditions on the distributions of predictor variables, conditions that are somewhat difficult to interpret and that appear to have been violated in some of our simulations. Similarly, even outside of the generalized linear model framework the AUC approach has an intuitive basis for yielding linear predictors with operating characteristics that are optimal amongst linear predictors while the logistic regression approach does not at this point. We have noted that the work of Eguchi and Copas (2002) may lead to insights into the apparently good performance of logistic regression but this avenue of reasoning needs to be developed much further.

A disadvantage of the AUC approach relative to logistic regression is that it is computationally more difficult to maximize  $\widehat{\text{AUC}}(b)$ . The objective function is not smooth, so that optimization procedures based on derivatives cannot be applied directly. Pepe and Thompson (2000) have previously suggested using the  $\widehat{\text{AUC}}(\cdot)$  function to find linear predictors (although they did not recognize connections with the generalized linear model or optimality of the resulting decision rules). They suggest numerical procedures to find  $\widehat{\beta}^{\text{AUC}} = \text{argmax}(\widehat{\beta}^{\text{AUC}}(b))$  that are based on modeling  $\text{AUC}(b)$  as a smooth function of  $b$ . They also discuss computational approaches when  $P > 2$ , i.e., when  $(\beta_2, \dots, \beta_P)$  is multidimensional.

Our exposition here deals with objective functions used to combine a given set of predictors into a linear combination. We have not dealt with the use of such functions to select amongst potential predictors or to assist in model formulation. Those avenues of research could be pursued. We also note that care must be taken in assessing the performance of the linear predictor function estimated from a dataset. If the same dataset is used to calculate the ROC curve for the estimated linear predictor function, then adjustments must be made that account for the potentially optimistic bias

of the performance of the predictor in the dataset from which it is derived. Work by Copas and Corbett (2002) gives direction on how to proceed.

As a secondary message from this paper we suggest that modeling of the risk function (or linear predictor) need only be done over a subset of the predictor space. The relevant subspace is identified by considering what ranges of operating characteristics are potentially useful for the corresponding decision rules. McIntosh and Pepe (2002) suggested restrictions to the predictor space based on considering the acceptable range of false-positive fractions. Here, we propose restrictions that include consideration of desirable true-positive fractions and suggest estimating the predictor coefficients by maximizing the area under the ROC curve that corresponds to the restricted ranges of true and false positive fractions. In practice, this idea is probably most useful when large amounts of data are available. An alternative approach is to allow some flexible modeling of predictor variables over the whole predictor space in order to relax linearity assumptions.

## 8. ACKNOWLEDGEMENTS

We would like to thank Noelle Noble and Gary Longton for assistance with preparing the manuscript, and Holly Janes for helpful comments on an earlier version of the paper.

## 8. REFERENCES

- Baker, S.G. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention *Biometrics* **56**, 1082–1087.
- Copas, J.B. and Corbett, P (2002). Overestimation of the receiver operating characteristic curve for logistic regression *Biometrika* **89**(2), 315–331.
- Eguchi, S. and Copas, J.B. (2002). A class of logistic-type discriminant functions *Biometrika* **89**(1), 1–22.

- Green, D.M. and Swets, J.A. (1966). *Signal detection theory and psychophysics*. Wiley, New York.
- Han A.K. (1987). Non-parametric analysis of a generalized regression model. The maximum rank correlation estimator. *Journal of Economics* **35**, 303–316
- Li, K-C. and Duan, N. (1989). Regression analysis under link violation. *The Annals of Statistics* **17**, 1009–1052.
- McIntosh M.S., Pepe M.S. (2002) Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657–664.
- Neyman, J. and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London, Series A* **231**, 289–337.
- Pepe, M.S. (2003) *The Statistical Evaluation of Medical Tests for Classification and Prediction* Oxford University Press, United Kingdom.
- Pepe, M.S. and Thompson M.L. (2000) Combining diagnostic test results to increase accuracy. *Biostatistics* **1**(2) 123–140.
- Pepe M.S., Etzioni R., Feng Z., Potter J.D., Thompson M., Thornquist M., Winget M., Yasui Y. (2001) Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**(14), 1054–1061.
- Prentice, R.L. and Pyke, R. (1979) Logistic disease incidence models and case-control studies. *Biometrika* **66** 403–412.
- Sherman R.P. (1993) The limiting distribution of the maximum rank correlation estimator. *Econometrics* **61**, 123–137.
- Zhou, X-H., Obuchowski, N.A., McClish, D.K. (2002) *Statistical Methods in Diagnostic Medi-*

*cine* Wiley, New York.

Table 1: Mean and sampling standard deviation (SSD) of  $\hat{\beta}_2$  and  $\text{AUC}(\hat{\beta}_2)$  across simulations from both the maximum logistic likelihood approach ( $\hat{\beta}_2^L$ ) and the maxAUC approach ( $\hat{\beta}_2^{\text{AUC}}$ ). Results are based on 1000 simulated datasets for each sample size. The true value of  $\beta_2 = 0.5$  and the optimal linear predictor has  $\text{AUC}(\beta_2) = 0.808$ .

		$(n_D, n_{\bar{D}})$	(100,100)		(100, 1000)		(1000,100)	
			Mean	SSD	Mean	SSD	Mean	SSD
$\hat{\beta}_2$	Logistic mle		.522	.138	.504	.082	.505	.090
	AUC		.528	.171	.511	.105	.507	.100
$\text{AUC}(\hat{\beta}_2)$	Logistic mle		.804	.0034	.802	.0014	.804	.0016
	AUC		.805	.0047	.802	.0021	.804	.0019

Table 2: Mean and sampling standard deviation of  $\widehat{\beta}_2$  and  $\text{AUC}(\widehat{\beta}_2)$  across simulations from the misspecified link function models, when estimation is based on the logistic likelihood and on the empirical AUC. Results are from 1000 simulated datasets for each sample size. The true value of  $\beta_2 = 0.5$  and the optimal linear predictor has  $\text{AUC}(\beta_2) = 0.933$  when  $H(x) = H_1(x)$  and  $\text{AUC}(\beta_2) = 0.937$  when  $H(x) = H_2(x)$ .

		$(n_D, n_{\bar{D}})$	(100,100)		(100, 1000)		(1000,100)	
			Mean	SSD ( $\times 10$ )	Mean	SSD ( $\times 10$ )	Mean	SSD( $\times 10$ )
$H_1(x)$	$\widehat{\beta}_2$	Logistic mle	.500	.699	.504	.403	.500	.422
		AUC	.503	.824	.501	.562	.506	.571
	$\text{AUC}(\widehat{\beta}_2)$	Logistic mle	.930	.016	.928	.007	.931	.007
		AUC	.930	.021	.928	.011	.931	.010
$H_2(x)$	$\widehat{\beta}_2$	Logistic mle	.581	.146	.552	.062	.610	.135
		AUC	.525	.157	.515	.121	.515	.098
	$\text{AUC}(\widehat{\beta}_2)$	Logistic mle	.936	.0033	.935	.0009	.936	.0032
		AUC	.937	.0041	.937	.0027	.937	.0018

Table 3: Sample mean and sampling standard deviations of  $\hat{\beta}_2$  and  $\text{AUC}(\hat{\beta}_2)$  from the misspecified linear predictor model.

		$(n_D, n_{\bar{D}})$		(100,100)		(100, 1000)		(1000,100)	
				Mean	SSD	Mean	SSD	Mean	SSD
$\hat{\beta}_2$	Logistic mle			.463	.153	.462	.0615	.411	.104
	AUC			.466	.184	.440	.0942	.446	.115
$\text{AUC}(\hat{\beta}_2)$	Logistic mle			.877	.0097	.883	.0049	.886	.0066
	AUC			.876	.0116	.884	.0068	.884	.0082

Table 4: Sample mean and sampling standard deviation of  $\hat{\beta}_2$  and  $\text{AUC}(\hat{\beta}_2)$  from the normal mixture model.

		$(n_D, n_{\bar{D}})$	(100,100)		(100, 1000)		(1000,100)	
			Mean	SSD	Mean	SSD	Mean	SSD
$\hat{\beta}_2$	Logistic mle		.483	.101	.445	.116	.593	.0387
	AUC		.447	.129	.447	.105	.446	.0832
$\text{AUC}(\hat{\beta}_2)$	Logistic mle		.89	.0024	.895	.0030	.895	.0015
	AUC		.89	.0037	.896	.0025	.896	.0022

Table 5: Sample mean and sampling standard deviation of  $\hat{\beta}_2$  and  $\text{rpAUC}(\hat{\beta}_2)$  from simulated data of the normal mixture model. Shown in the top two rows are results for estimates of  $\hat{\beta}_2$  using the restricted region logistic likelihood (rLogistic) and the restricted region AUC (rAUC). In the lower four rows corresponding values of  $\text{rpAUC}(\hat{\beta}_2)$  are shown for both methods as well as those using the linearly assumption over the entire predictor space (logistic mle and AUC).

	$(n_D, n_{\bar{D}})$	(100,100)		(100, 1000)		(1000,100)	
Method		Mean	SSD	Mean	SSD	Mean	SSD
rLogistic		.280	.158	.281	.159	.308	.087
rAUC		.174	.549	.031	.229	.258	.699
rLogistic		.786	.027	.796	.028	.785	.016
rAUC		.812	.026	.835	.011	.812	.026
Logistic mle		.748	.014	.761	.021	.748	.0056
AUC		.755	.021	.760	.019	.774	.0163

Figure 1: The ROC curve associated with the risk score,  $P(D = 1|Y_1, Y_2)$ , for the simulation settings presented in Tables 1, 2 and 3. The curves are labeled 1 (setting of Section 3.1), 2 (Section 3.2 with  $g = L(H - 1(x))$ ), 3 (Section 3.2 with  $g = L(H_2(x))$ ) and 4 (Section 3.3 equation (6)).

**Figure 1**

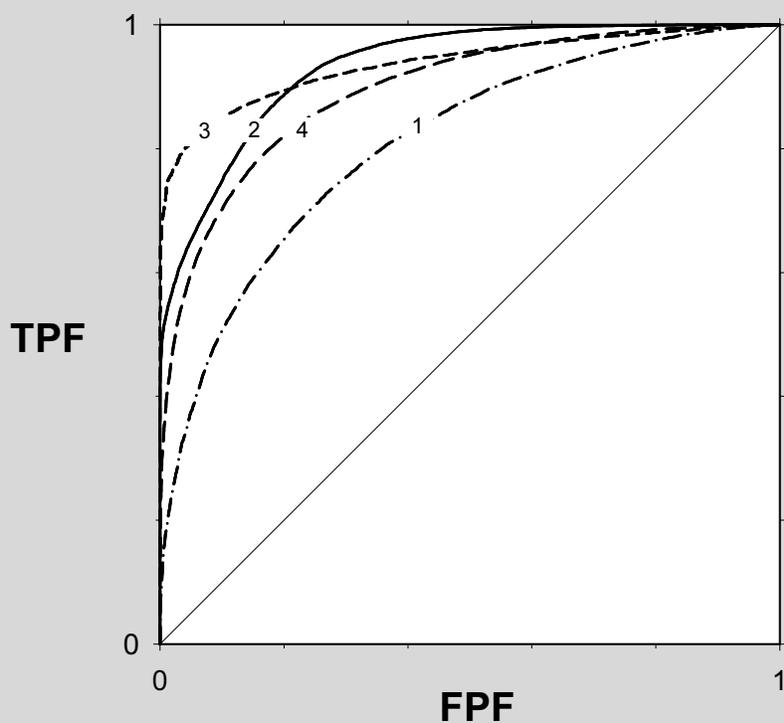


Figure 2: Plot of the function  $\text{logit}\{L(H(\cdot))\}$  over the range of simulated predictors corresponding to the 1<sup>st</sup> to 99<sup>th</sup> percentiles in controls. The link function  $L\{H(\cdot)\}$  is not approximated by a logistic form.

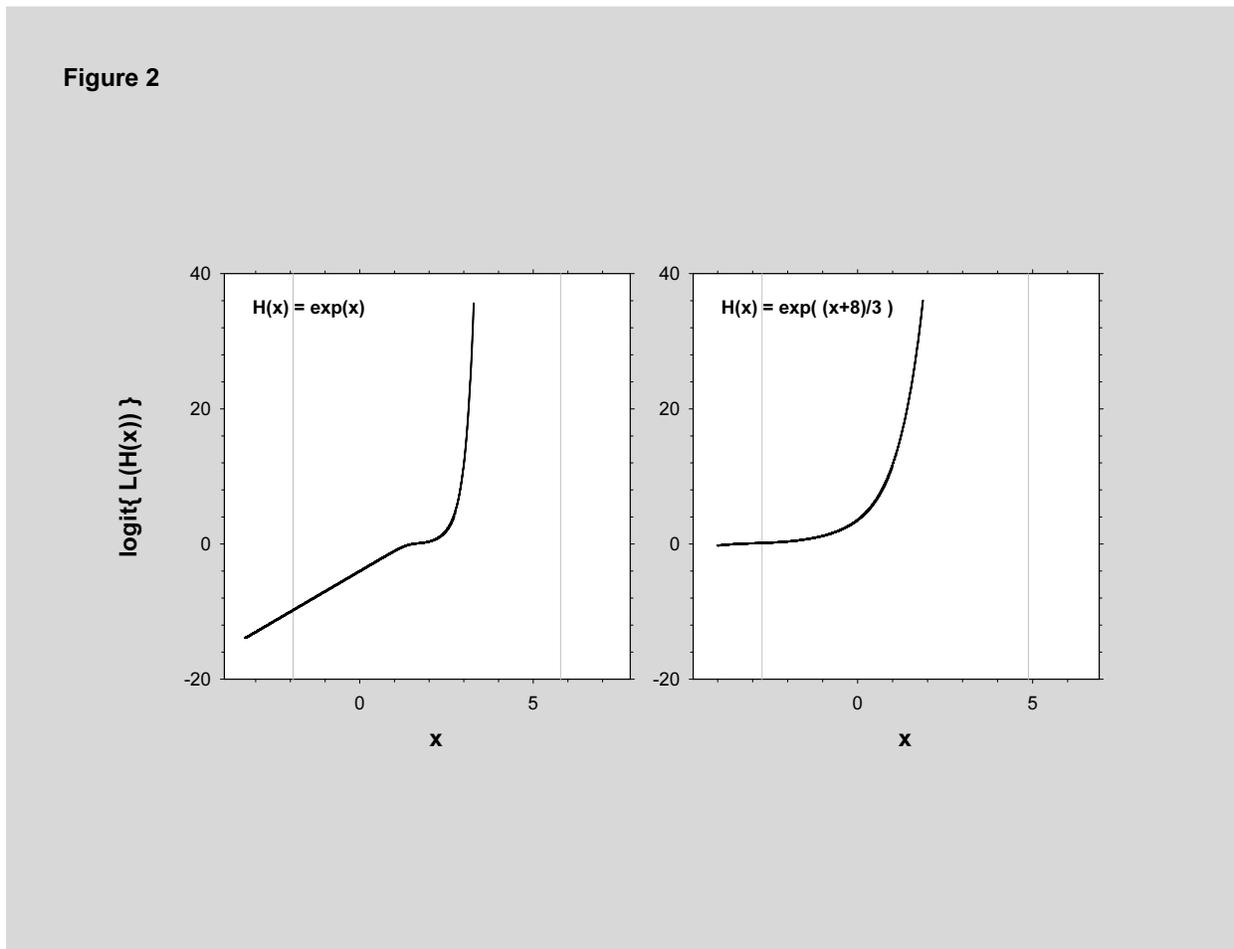


Figure 3: Schematic diagrams of (A) the restricted region of the ROC space and (B) the corresponding ranges of the risk score. The point  $q_{\bar{D}}(f_0)$  is the  $1 - f_0$  quantile of  $RS(Y)$  in the population  $D = 0$  and  $q_D(t_0)$  is the  $(1 - t_0)$  quantile in the population  $D = 1$ .

