INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD • INDIA

Research and Publications

ROC Curve Analysis for Randomly Selected Patients

Tathagata Bandyopadhyay
Sumanta Adhya
Apratim Guha

W.P. No. 2015-07-02
July 2015

The main objective of the working paper series of the IIMA is to help faculty members, research staff and doctoral students to speedily share their research findings with professional colleagues and test their research findings at the pre-publication stage. IIMA is committed to maintain academic freedom. The opinion(s), view(s) and conclusion(s) expressed in the working paper are those of the authors and not that of IIMA.

INDIAN INSTITUTE OF MANAGEMENT
AHMEDABAD-380 015
INDIA

ROC Curve Analysis for Randomly Selected Patients

Tathagata Bandyopadhyay[1], Sumanta Adhya[2] and Apratim Guha[3]

Abstract

Receiver operating characteristic (ROC) curves and the area under the curve (AUC) are widely used in medical studies to examine the effectiveness of markers in diagnosing diseases. In most of the existing literature for ROC curve analysis it is assumed that the healthy and the diseased populations are independent of each other, which may lead to bias in the studies. In this paper we consider the disease status as a binary random variable. Assuming the disease status is determined by a latent variable and the marker and the latent variable have a bivariate normal distribution, we derive the properties of the ROC curve and the AUC. We also look at the problem of choosing optimum combination of markers when multiple markers are present. Limiting distributions are obtained and confidence intervals are discussed as well. A small simulation study is performed which confirms the superiority of our methods over the general practice of considering the two populations to be independent.

1 Introduction

ROC curves and in particular AUC are widely used in medical studies to examine the effectiveness of markers used to diagnose diseases. Consider a study where for each individual disease status, observations on a continuous marker and a set of related covariates are available. Let us assume that for an individual the higher the value of the marker more is the chance of being diseased. In determining the ROC curves it is almost always a priori assumed the samples for the diseased (case) and healthy (control) individuals are obtained from two different, independent, populations. To be precise, let Y_1 and Y_2 respectively denote the random variables representing marker values for the healthy and diseased individuals, which are distributed independently. For a given value $c \in \mathbb{R}$ we define the true positive rate at c $(TPR(c))$ to be equal to $P(Y_2>c)$ and the false positive rate at c $(FPR(c))$ to be equal to $P(Y_1>c)$. The ROC curve is then obtained by joining the points $(FPR(c), TPR(c))$ for c varying over the real line, see Pepe et

[1] P & QM Area, Indian Institute of Management, Ahmedabad, India. Email: tathagata@iimahd.ernet.in

[2] Department of Statistics, West Bengal State University, Barasat, India. Email: sumanta.adhya@gmail.com

[3] P & QM Area, Indian Institute Of Management, Ahmedabad, India. Email: apratim@iimahd.ernet.in

al. **(**2001) and Baker (2003). It can be mathematically shown that AUC = $P(Y_2>Y_1)$, see Bamber (1975).

The setup mentioned above is the common practice but always not very sensible. Consider the situation, for example, when one wishes to evaluate the performance of a biomarker risk prediction, as considered by Pepe et al. (2012), Kerr and Pepe (2011) and the references within. When one considers these two groups to be independent, one faces the dilemma whether to "match" the biomarkers for case and control groups or not. The common practice of matching is problematic, see Janes and Pepe (2008), but not matching may also lead to bias. There is perhaps a need for a different approach.

To eliminate the bias that arises from assuming the diseased and healthy individuals to be from two different populations, one option is to consider them to be from the same population but at two different "states", namely: "diseased" or "healthy". It may easily be achieved by assuming the disease status to be a dichotomous random variable instead, as in Pepe et al. (2012). In that case we may say that we have a random sample of size n from the study population. For each selected individual we then observe the disease status (D) which is binary, (where $D = 1$ means diseased and $D = 0$ means healthy), the value of the marker (Y), and the covariates (x). Assuming an underlying latent variable Y_D for the binary variable D such that $Y_D > 0 \Leftrightarrow D = 1$ (diseased) and $Y_D < 0 \Leftrightarrow D = 0$ (healthy) and a bivariate absolutely continuous distribution for (Y,Y_D) we may carry out ROC analysis adjusting for the effect of covariates x. One may easily extend this analysis to the case where more than one marker are available by looking at an appropriate multivariate structure. In this paper we restrict our analysis to a bivariate/ multivariate normal structure for convenience. In case of departure from normality one may think of appropriate transformations, see Schisterman et al. (2004) for motivation.

In Section 2, we introduce the model and find an expression for the area under ROC curve (AUC) for any given covariate level x_0. In Section 3 we find the maximum likelihood estimates (MLE) of the model parameters. In Section 4 the large sample distributions of the ROC values and AUC are established, and an interval estimate of AUC given the covariate value (level) x_0 has been obtained using bootstrap. In Section 5, a simulation study has been carried out to compare the ROC curve obtained by our approach with that of the standard approach. Concluding discussions are given in Section 6.

2 A Parametric Model

In the standard ROC curve analysis with marker Y following a normal distribution it is assumed that the parameters of the distributions of Y for diseased and healthy patients are independent, see Pepe (2003). As already established in Section 1, this assumption seems to be restrictive as it may require matching the biomarkers. In this paper we propose a simple parametric model obtained from the joint distribution of the binary variable indicating presence/absence of disease, a set of markers assumed to be continuous, and a set of covariates, also assumed to be continuous. For some useful references in analysis of ROC curve when covariates are present, one may refer to Janes et al. (2009), Pardo-Fernández et al. (2014) and the references within. Note that our approach could be adapted to situations where some or all of the markers and/or the covariates may be categorical, as well as to the complex situations where nonparametric or semiparametric models are required. We would discuss it in the sequels.

To begin with suppose we have a single continuous marker Y for discriminating between the diseased and the healthy individuals. Although it is possible to extend our analysis to multiple markers, in that case the problem of choosing an optimal marker combination arises, which is not of interest for our current work. However, we briefly discuss the application of our methodology in such a situation in Section 6.

Suppose we have a random sample of size n from the study population and the data are in the form $(y_i, d_i, x_i), i = 1,\cdots,n.$. As mentioned in Section 1, we assume an underlying latent variable Y_D for D. We now assume that (Y,Y_D) has a bivariate normal distribution with respective means $\mu_Y(x_1) = \beta_1^T x_1$ and $\mu_{Y_D}(x_2) = \beta_2^T x_2$, respective variances σ_Y^2 and 1 and correlation coefficient ρ. Note that x_1 and x_2 are subsets of x. Thus the distribution of the marker Y given $Y_D > 0$ $(Y_D < 0)$) represents the distribution of Y for the diseased and healthy patients respectively. Writing the model parameters $\theta = (\beta_1^T, \beta_2^T, \sigma_Y^2, 1)^T$ the conditional distribution of Y given $Y_D > 0$ and $Y_D < 0$ are respectively given by:

$$\begin{aligned} f_D(y\,|\,x) &= f(y\,|\,x, Y_D > 0) = f_Y(y)a_D(x,y;\theta) \\ f_{\bar{D}}(y\,|\,x) &= f(y\,|\,x, Y_D < 0) = f_Y(y)a_{\bar{D}}(x,y;\theta) \end{aligned} \quad \ldots (2.1)$$

where

$$a_D(x,y;\theta) = \Phi\left(\left(1-\rho^2\right)^{-1/2}\mu_{Y_D|Y}\right)\Big/\Phi\left(\beta_2^T x_2\right),$$

$$a_{\bar{D}}(x,y;\theta) = \Phi\left(-\left(1-\rho^2\right)^{-1/2}\mu_{Y_D|Y}\right)\Big/\Phi\left(-\beta_2^T x_2\right),$$

$$\mu_{Y_D|Y} = \beta_2^T x_2 + \frac{\rho}{\sigma_Y}\left(y - \beta_1^T x_1\right)$$

and $\Phi(.)$ is the cdf of the standard normal distribution.

We have assumed at the outset that $0 \le \rho \le 1$. For $0 < \rho < 1$, the distributions $f_D(y\,|\,x)$ and $f_{\bar{D}}(y\,|\,x)$ are not symmetric. Also $a_D(x,y;\theta)$ ($a_{\bar{D}}(x,y;\theta)$) increases (decreases) in y from 0 $\left(1\big/\Phi\left(-\beta_2^T x_2\right)\right)$ to $1\big/\Phi\left(\beta_2^T x_2\right)$(0). Note that for y greater (less) than $\beta_1^T x_1 - \frac{\sigma_Y}{\rho}\left(1-\left(1-\rho^2\right)^{1/2}\right)\beta_2^T x_2 (= y_0(x;\theta)$, say), $a_D(x,y;\theta)$($a_{\bar{D}}(x,y;\theta)$) is greater (less) than unity. Thus $f_D(y\,|\,x)$ becomes greater (less) than $f_Y(y)$ for y greater (less) than $y_0(x;\theta)$. The relation between $f_{\bar{D}}(y\,|\,x)$ and $f_Y(y)$ is exactly the opposite. In other words, $f_D(y\,|\,x)$ becomes positively skewed and $f_{\bar{D}}(y\,|\,x)$ negatively skewed. The amount of inflicted skewness depends on the value of ρ. The closer it is to 1, the skewness becomes more pronounced. In the extreme cases viz. $\rho = 0$ and $\rho = 1$ the distributions becomes identical and completely separated.

We now state and prove the following result.

Theorem 1: *For $0 < \rho < 1$, the distribution of $f_D(y\,|\,x)$ is stochastically larger than the distribution of $f_{\bar{D}}(y\,|\,x)$.*

Proof: Note that $f_D(y\,|\,x) < f_Y(y)$ ($f_{\bar{D}}(y\,|\,x) > f_Y(y)$) for $y < y_0(x;\theta)$ and $f_D(y\,|\,x) > f_Y(y)$ ($f_{\bar{D}}(y\,|\,x) < f_Y(y)$) for $y > y_0(x;\theta)$.

Thus for $y < y_0(x;\theta)$, we get

$$F_D(y\,|\,x) = \int_{-\infty}^{y} f_D(t\,|\,x)dt < \int_{-\infty}^{y} f_Y(t)dt = \int_{-\infty}^{y} f_{\bar{D}}(t\,|\,x)dt = F_{\bar{D}}(y\,|\,x)\,.$$

Similarly, for $y > y_0(x;\theta)$, $\bar{F}_D(y\,|\,x) > \bar{F}_D(y\,|\,x) \Leftrightarrow F_D(y\,|\,x) < F_{\bar{D}}(y\,|\,x)$, which completes the proof.

To draw the ROC curve we need to find True positive rate (TPR) $\bar{F}_D(c)\,(=1-F_D(c))$ and False Positive Rate (FPR) $\bar{F}_{\bar{D}}(c)\,(=1-F_{\bar{D}}(c))$ for any real c and then to plot $(\bar{F}_{\bar{D}}(c), \bar{F}_D(c))$ for different values of c for a given x. In fact, ROC curve is plot of the function $\bar{F}_D\left(\bar{F}_{\bar{D}}^{*-1}(t)\right)$ for $t \in (0,1)$ for a given x where $\bar{F}_{\bar{D}}^{*}(c)$ is a survival function of the density $\varphi(z)\Phi\left(-\left(1-\rho^2\right)^{1/2}\left(\rho z + \beta_2^T x_2\right)\right)\Phi^{-1}\left(\beta_2^T x_2\right)$, $z \in$ □, where $\phi(.)$ and $\Phi(.)$ denote the density and the distribution function of the standard normal distribution. This is often called the adjusted ROC curve, see Janes et al. (2009).

The area under the ROC curve (AUC) is an important measure of diagnostic accuracy**.** For examples, properties and applications see Metz *et al.* (1984), Su and Liu (1993), Zhou *et al.* (2002), Liu *et al.* (2005), Ma and Huang (2005) and Wang *et al.* (2007). For a given x_0, the AUC is given by $\mathrm{AUC}_{x_0} = P\left(Y_1 > Y_2\right)\left(= P\left(Y_1 > Y_2\,|\,x = x_0\right)\right)$ where Y_1 and Y_2 are independent random variables with probability density $f_D(.)$ and $f_{\bar{D}}(.)$ respectively. This may be referred to as “adjusted AUC”.

3 Maximum Likelihood Estimation

Among the various choices available for estimating the ROC curve, perhaps the most popular choice among the parametric methods is to employ the maximum likelihood estimation (MLE), which dates back to Ogilvie and Creelman (1968), also see Dorfman and Alf (1969) for further elaboration. This method, among other advantages, readily allows us to obtain asymptotic confidence bands for the ROC curve and also helps to estimate the AUC, as we explore in Section 3.1.

Let us write

$$\theta = \left(\beta_1^T, \beta_2^T, \sigma_Y^2, \rho\right)^T, \; v_i(\theta) = \left(1-\rho^2\right)^{-1/2}\left(\beta_2^T x_{2i} + \frac{\rho}{\sigma_Y}\left(y_i - \beta_1^T x_i\right)\right),$$

and

$$w_i(\theta) = \varphi\left(v_i(\theta)\right) \Big/ \left[\Phi\left(v_i(\theta)\right)\left(1-\Phi\left(v_i(\theta)\right)\right)\right].$$

Then the full log likelihood for n observations is given by

$$l(\theta) = \sum_i ln\, f_Y\left(y_i \mid x_{1i}; \beta_1, \sigma_Y^2\right) + \sum_i d_i\, ln\left(\Phi\left(v_i(\theta)\right)\right) + \sum_i (1-d_i)\, ln\left(-\Phi\left(v_i(\theta)\right)\right)$$

where $f_Y\left(y_i \mid x_{1i}; \beta_1, \sigma_Y^2\right)$ is the density of Y_i given x_i.

The corresponding score equations are:

$$\partial l/\partial\beta_1 = \sigma_Y^{-2}\sum_i x_{1i}\left(y_i - x_{1i}^T\beta_1\right) - \rho\sigma_Y^{-1}\left(1-\rho^2\right)^{-1/2}\sum_i x_{1i} w_i(\theta)\left(d_i - \Phi\left(v_i(\theta)\right)\right);$$

$$\partial l/\partial\beta_2 = \left(1-\rho^2\right)^{-1/2}\sum_i x_{2i} w_i(\theta)\left(d_i - \Phi\left(v_i(\theta)\right)\right);$$

$$\partial l/\partial\sigma_Y = -n\sigma_Y^{-1} + \sigma_Y^{-3}\sum_i\left(y_i - x_{1i}^T\beta_1\right)^2 - \rho\sigma_Y^{-2}\left(1-\rho^2\right)^{-1/2}\sum_i\left(y_i - x_{1i}^T\beta_1\right) w_i(\theta)\left(d_i - \Phi\left(v_i(\theta)\right)\right)$$

and

$$\partial l/\partial\rho = \sigma_Y^{-1}\left(1-\rho^2\right)^{-3/2}\sum_i\left(y_i - x_{1i}^T\beta_1\right)w_i(\theta)\left(d_i - \Phi\left(v_i(\theta)\right)\right).$$

We can see that no closed form solution is available, and it is a bit complicated to obtain the information matrix. However, one can simply obtain a numerical solution to the likelihood using some statistical package. As for example, one may use the '*optim*' package available with the statistical software *R*, which we use in the simulation exercise carried out in Section 5.

3.1 AUC at the Maximum Likelihood

We now discuss how to employ the maximum likelihood estimate obtained above to estimate the adjusted AUC.

Recall from Section 2.1 that the adjusted AUC at a given *x* is given by

$$\mathrm{AUC}_x = H(x;\theta) = -\int \overline{F}_D(y)d\overline{F}_{\overline{D}}(y), \qquad \ldots(3.3)$$

where $\overline{F}_D(y) = \int_y^\infty f_D(u)du$ and $\overline{F}_{\overline{D}}(y) = \int_y^\infty f_{\overline{D}}(u)du$ are the survival functions of the conditional distributions of *Y* given *D* and $\overline{D}$ respectively.

Now, by direct plug-in of the MLE, we can write the estimated AUC at x as

$$\mathrm{AUC}_x = H\left(x;\hat{\theta}\right) \qquad \ldots(3.4)$$

which is trivially a consistent estimator of AUC_x under the present setup. In Section 4.1 we establish the asymptotic normality of the above estimator and develop a confidence interval for the adjusted AUC.

4 Asymptotic Properties of the ROC Curve and the AUC

In this section we establish the asymptotic normality of the ROC curve and the AUC for the given setup at the parametric rate using the asymptotic properties of the MLE $\hat{\theta}$ of

$\theta = \left(\beta_1^T, \beta_2^T, \sigma_Y^2, \rho\right)^T$, and hence discuss the development of asymptotic confidence intervals for the same.

4.1 Asymptotic Normality of the ROC Curve

It can easily be shown that the MLE $\hat{\theta}$ satisfies the conditions of Theorem 7.5.1 of Lehmann (1999), so that the following hold true

(a) $$\hat{\theta} - \theta = O_P\left(n^{-1/2}\right) \qquad \ldots(4.1)$$

and

(b) $$n^{1/2}\left(\hat{\theta} - \theta\right) \to N\left(0, \bar{I}\left(\theta\right)\right), \qquad \ldots(4.2)$$

where

$$\bar{I}(\theta) = \lim_{n\to\infty} n^{-1} E\left\{-\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}\right\}.$$

The population ROC curve is obtained by plotting $\mathrm{ROC}_x(t) = \bar{F}_D\left(\bar{F}_{\bar{D}}^{-1}\left(t\right)\right)$ against $t \in (0,1)$.

The natural estimate of $ROC_x(t)$ is $\mathrm{ROC}x(t) = \hat{\bar{F}}_D\left(\hat{\bar{F}}_{\bar{D}}^{-1}(t)\right)$, where $\hat{\bar{F}}_D$ and $\hat{\bar{F}}_{\bar{D}}$ are the values of $\bar{F}_D$ and $\bar{F}_{\bar{D}}$ respectively, obtained by plugging in $\hat{\theta}$ for θ.

***Theorem 2**: Given $t \in (0,1)$ and $\theta \in \square$, as $n \to \infty$,*

$$n^{1/2}\left\{\mathrm{ROC}x\left(t\right) - \mathrm{ROC}_x\left(t\right)\right\} \xrightarrow{d} N\left(0, v_x\left(t;\theta\right)\right)$$

where the asymptotic variance $v_x(t;\theta)$ is given by

$$v_x(t;\theta) = h_x^T(t;\theta) I(\theta)^{-1} h_x(t;\theta),$$

and

$$h_x(t;\theta)=E_D\left\{\left(\frac{\partial \ln f_D(Y\,|\,x)}{\partial\theta}\right)I(Y>t)\right\}-\frac{f_D(t\,|\,x)}{f_{\bar{D}}(t\,|\,x)}E_{\bar{D}}\left\{\left(\frac{\partial \ln f_{\bar{D}}(Y\,|\,x)}{\partial\theta}\right)I(Y>t)\right\}$$

where $I(\theta)$ *is the Fisher's Information matrix and* $E_D(.)$ *and* $E_{\bar{D}}(.)$ *are expectation with respect to distributions* $\{Y\,|\,x, D=1\}$ *and* $\{Y\,|\,x, D=0\}$ *respectively. Further, a consistent estimator of the asymptotic variance is given by*

$$\hat{v}_x\left(t;\hat{\theta}\right)=\hat{h}_x^T\left(t;\hat{\theta}\right)\hat{I}\left(\hat{\theta}\right)^{-1}\hat{h}_x\left(t;\hat{\theta}\right)$$

where

$$\hat{h}_x(t;\theta)=n^{-1}\sum_{i=1}^{n}\frac{\partial \ln f_D(Y_i\,|\,x)}{\partial\theta}I(Y_i>t,D_i=1)-c_x(t;\theta)n^{-1}\sum_{i=1}^{n}\frac{\partial \ln f_{\bar{D}}(Y_i\,|\,x)}{\partial\theta}I(Y_i>t,D_i=0)$$

for some suitably chosen $c_x(t,\theta)$.

Proof. Let $c_x(t;\theta)=f_D(t\,|\,x)/f_{\bar{D}}(t\,|\,x)$. Using the delta method as in the proof Theorem 1 (*ii*) of *Ma et al.* (2010), we have

$$\mathrm{ROC}_x(t)-\mathrm{ROC}_x(t)=\left\{\hat{\bar{F}}_D(t)-c_x(t;\theta)\hat{\bar{F}}_{\bar{D}}(t)\right\}-\left\{\bar{F}_D(t)-c_x\left(t;\theta\right)\bar{F}_{\bar{D}}(t)\right\}.$$

Further applying a Taylor series expansion about θ, we get

$$n^{1/2}\left(\mathrm{ROC}_x\left(t\right)-\mathrm{ROC}_x\left(t\right)\right)=n^{1/2}\left(\hat{\theta}-\theta\right)^T h_x\left(t;\theta\right)+O_P\left(n^{-1/2}\right)$$

and hence the asymptotic normality is established.

4.2 Large Sample CI for the AUC

We have seen in Section 3.1 that AUC$_x$ = $H(x;\theta)$ is a continuously derivable function of θ. Hence, using the delta method, from (4.1) and (4.2) above, we have

$$n^{1/2}\{H(x;\hat{\theta})-H(x;\theta)\}\to N(0,\sigma_H^2(x;\theta)),\ \sigma_H^2(x;\theta)=H'_\theta(x;\theta)^T\,\bar{I}(\theta)^{-1}H'_\theta(x;\theta).$$

In view of the above, theoretically we have the means to obtain the asymptotic variance $\sigma_H^2(x;\theta)$, and hence obtain confidence intervals for AUC. Unfortunately, although computation can be further simplified by replacing $\bar{I}(\theta)$ with limiting average observed information $\bar{I}_o(\theta)=-\lim_{n\to\infty} n^{-1}\partial^2 l(\theta)/\partial\theta\partial\theta^T$, in general computation of $H'_\theta(x;\theta)$ will be complicated. In this article we do not obtain a confidence interval for the AUC, but see Section 6.3 for a possible solution.

5 Some Simulation Studies

In this section we look at the performance of the proposed methodology through some simulation studies. We compare the performance of the proposed method with the existing method of obtaining the ROC curves and AUC considering the diseased and the healthy subjects to be part of two independent populations.

We perform the simulation exercises by first sampling the covariate values (x) from mixture normal distributions, and then generating the marker and latent variable values for disease status from some bivariate normal distributions.

Step 1: Draw a random sample $x_1,...,x_n$ from a mixture normal density. Let it be $X_n=(x_1,...,x_n)$. Then for each x generate (y,y_d) value using bivariate normal density with means $\mu_{y|x}=\alpha_o+\alpha_1 x$ and $\mu_{y_d|x}=\beta_0+\beta_1 x$, variances σ_y^2 and 1, and correlation ρ. Here y is the marker and y_d is the latent variable for disease or healthy unit. Define $d_i=1$ (disease) if $y_{d_i}>0$ and 0 (healthy), otherwise, $i=1,...,n$. The sample is $\{(y_i,d_i,x_i):i=1,...,n\}$. Given X_n, we repeat above procedure to obtain $B(=100)$ independent set of samples $S_b=\{(y_{ib},d_{ib},x_i):i=1,...,n\}$, $b=1,...,B$.

Step 2: Consider two groups: the first group is represented by H (for healthy, $d_{ib}=0$) and the second group is represented by D (for diseased, $d_{ib}=1$). Thus the b-th sample is split into two subsamples: $S_b^H = \{(y_{ib}, x_i) : d_{ib} = 0, i = 1,\ldots,n\}$ assigned to group H (with sample size $n_b^H = n - \sum_i d_{ib}$) and $S_b^D = \{(y_{ib}, x_i) : d_{ib} = 1, i = 1,\ldots,n\}$ assigned to group D (with sample size $n_b^H = \sum_i d_{ib}$), $b = 1,\ldots,B$. Then fit linear regressions based on samples S_b^H and S_b^D and estimate $\mathrm{FPR}_{b,x}^1$ and $\mathrm{TPR}_{b,x}^1$ for the set of different threshold values T and the x-values are: $x^1 = \min(X_n)$, $x^2 =$ first quartile of X_n, $x^3 =$ 2nd quartile of X_n, $x^4 =$ third quartile of X_n, and $x^5 = \max(X_n)$. Finally consider set of t points ($\mathrm{FPR}_x^1 = B^{-1}\sum_b \mathrm{FPR}_{b,x}^1$, $\mathrm{TPR}_x^1 = B^{-1}\sum_b \mathrm{TPR}_{b,x}^1$) for each choice of x, $x = x^1,\ldots,x^5$. The set of points is represented by ROC_x^1, $x = x^1,\ldots,x^5$. The adjusted AUC is then obtained through numerical integration.

Step 3. In Step 3 we compute the ROC, ROC_x^2, $x = x^1,\ldots,x^5$, using proposed methodology. Here $FPR_{b,x}^1$ and $TPR_{b,x}^1$ is evaluated from the fitted bivariate normal population consider in Step 1 based on the sample S_b. The rest is same as step 2. $x = x^1,\ldots,x^5$.

Step 4: Finally in Step 4 we plot (see Figures 1-2) two set of points (ROC_x^1 and ROC_x^2) for a fixed x-value in same graph to obtain covariate adjusted ROC curves for standard (step 2) and proposed (step 3) methodologies, $x = x^1,\ldots,x^5$. In Figures 1-4 the ROC curves corresponding to the standard method are represented by the red lines and the ROC curves corresponding to the proposed method are represented by the blue lines.

We perform the simulation exercises using two different models as described below.

Simulation 1: The marginal distribution of x is mixture of normal densities $N\left(1,(0.3)^2\right)$ and $N\left(-1,(0.3)^2\right)$ with equal probability. The joint distribution of (y, y_d) given x is bivariate normal with means $\mu_{y|x} = \alpha_o$ and $\mu_{y_d|x} = \beta_0 + \beta_1 x$, variances σ_y^2 and 1, and correlation ρ where values of the parameters are: $\alpha_0 = 0.1$, $\alpha_1 = 0.25$, $\beta_0 = 0.01$, $\beta_1 = 1$, $\sigma_y = 0.5$ and $\rho = 0.7$. The set T consists of $t = 61$ equidistant points (distance = 0.1) from -3 to 3.

Simulation 2: The steps are exactly similar to Steps 1-4 of simulation 1 except for the distribution of x and joint distribution of (y, y_d). The marginal distribution of x is mixture of normal densities $N(1,(0.1)^2)$ and $N(-1,(0.1)^2)$ with equal probability and the joint distribution of (y, y_d) given x is bivariate normal with means $\mu_{y|x} = \alpha_o$ and $\mu_{y_d|x} = \beta_0 + \beta_1 x$, variances σ_y^2 and 1, and correlation ρ. The values of parameters are: $\alpha_0 = 0.9$, $\beta_0 = 0$, $\beta_1 = 1$, $\sigma_y = 1$ and $\rho = 0.8$.

The simulations were performed for sample size n =100.

5.1 Discussion of Simulation Results

Table 1: The average adjusted AUC values for the two simulated data sets with sample size 100.

Values of the Covariate	Average Adjusted AUC values for n = 100			
	Simulation 1		Simulation 2	
	Standard Method	Proposed Method	Standard Method	Proposed Method
Minimum	0.8540	0.8669	0.9004	0.9733
First Quartile	0.8329	0.8403	0.9120	0.9609
Second Quartile/ Median	0.7985	0.8156	0.9152	0.9803
Third Quartile	0.7838	0.8359	0.8963	0.9817
Maximum	0.7683	0.8574	0.8799	0.9782

The resulting average adjusted AUC values for both the simulations are summarized in Table 1, and the average ROC curves are presented in Figures 1 and 2 for Simulations 1 and 2 respectively. From Figure 1, we can see that in Simulation 1, the standard methodology is slightly worse than the proposed methodology for covariates $x=x^1$, x^2 *and* x^3, but the proposed methodology performs significantly better at the third quartile and the maximum. This finding is also confirmed by looking at the average adjusted AUC values for Simulation 1, as given in Table 1. It may further be noted that the ROC curve for the proposed methodology seems to stochastically dominate the standard methodology in this case.

Simulation 2 looks at a situation when the two normal distributions mixed to simulate *x* are better separated. We can see that both methods perform better in this case, as expected, as they report higher AUC values, but the proposed methodology seems to perform significantly better than the standard method, as is evident from Figure 2 as well as the average adjusted AUC values reported in the last two columns of Table 1.

6 Discussions and Further Extensions

In this article we have discussed how one may consider the diseased and the healthy subjects to be from the same population, instead of two different independent populations, and obtain the ROC curve and AUC in such a situation. We have discussed the maximum likelihood estimates and their asymptotic properties for the proposed method to obtain the ROC, and have demonstrated its relative advantage over the standard methodology through a simulation exercise. In this section we now discuss some challenges that one may encounter applying our proposed methodologies and their solutions. We also discuss an additional simulation example with a smaller sample size to demonstrate the advantage of our proposed methodology over the standard methodology.

6.1 Dealing with Multiple Markers

In this article all our discussions were based on the case when we have only a single marker. It is possible to encounter a situation where we have multiple continuous markers, which is very common in medical practice. Selecting the "best" linear combination of the markers in that case is a well-documented problem, for example one may refer to Schisterman et al. (2004), Liu et al. (2005). Lin et al. (2011), Wang and Chang (2011), Yu et al. (2011) and Chang (2013) to name a few. Schisterman et al. (2004) consider the problem of obtaining the best linear combination of the markers that would maximize AUC assuming the markers to be distributed as independent multivariate normal distributions for the population of diseased and healthy individuals conditionally given the value of the covariates. Such assumptions are restrictive, and might be hard to defend. In this work, as before, we proceed without this assumption of independence.

Suppose $Y\left(p\times 1\right)$ represents the vector of markers. As before suppose $f_D(y)$ and $f_{\bar{D}}(y)$ represent the distributions of Y for the diseased and the healthy individuals. (Note that from now on we suppress the dependence on x in the notations unless there is a scope for confusion.) We now have the following theorem under the assumption that the likelihood ratio

$\lambda(y)\ \left(= f_D(y)/f_{\bar{D}}(y)\right)$ is a monotone function of $s^*(y)$, a suitably chosen function of y. Actually in the present setup $s^*(y)$ is a linear function of the components of y which is to be defined later in this section.

Theorem 3: *The best combination of the markers in the sense of maximizing AUC is given by* $s^*(y)$.

Proof: Consider the testing of the simple null hypothesis that the distribution of the marker Y is $f_{\bar{D}}(y)$ against the simple alternative that the distribution of the marker Y is $f_D(y)$. Neyman-Pearson Lemma entails that the optimum test for the problem is given by $\left\{\lambda : \lambda(y) = f_D(y)/f_{\bar{D}}(y) > k\right\}$ where k is determined by the size condition.

Since $\lambda(y)$ is a monotone increasing function of $s^*(y)$, the optimum test can be equivalently written as $\left\{y : s^*(y) > c\right\}$. Further the test is optimum in the Neyman-Pearson sense, so it maximizes $P_D\left\{y : s(y) > c\right\}$ subject to the constraint that $P_{\bar{D}}\{y : s(y) > c\}$ is fixed for any real c where $s(y)$ is any other function of y.

For a given value of c the ROC curve corresponding to a combination of markers $s(y)$ plots TPR(c) $= P_D\{y : s(y) > c\}$ against FPR(c) $= P_{\bar{D}}\{y : s(y) > c\}$. From the above it is evident that the ROC curve corresponding to $s^*(y)$ maximizes the vertical height at any point on the horizontal axis and hence maximizes the AUC.

Note. Theorem 3 is a generalization of the result considered in Section 3 of Schisterman et al. (2004).

We will now apply the above theorem to our set-up to find the optimum combination of makers. Let us assume that $(Y^T(1\times p), Y_D)^T$ has a multivariate normal distribution where Y is the vector of marker values or an appropriate transformation of it. Then the mean vector and variance-covariance matrix of $(Y^T, Y_D^T)^T$ are respectively

$$\mu=\left(\mu_Y^T,\mu_{Y_D}\right)^T \text{ and } \Sigma=\begin{bmatrix}\Sigma_{YY}(p\times p) & \sigma_{YY_D}^T(1\times p)\\ \sigma_{YY_D}(p\times 1) & 1\end{bmatrix},$$

where

$$\mu_Y^T=\left(\beta_1^T x_1,\cdots,\beta_p^T x_p\right),\ \mu_{Y_D}=\beta_{p+1}^T x_{p+1},\ \Sigma_{YY}=(\sigma_{YY}) \text{ and } \sigma_{YY_D}=\left(\sigma_{11Y}^{1/2}\rho_{1d},\cdots,\sigma_{ppY}^{1/2}\rho_{pd}\right)^T$$

and Y_D, as before, represents the latent variable corresponding to the binary variable D defined in Section 2.

Writing $\theta=\left(\mu,\Sigma,\beta_1,....,\beta_{p+1}\right)$, $x=\left(x_1^T,\cdots,x_{p+1}^T\right)^T$, we thus have

$$f_D(y)=f_Y(y)a_D(y;x,\theta) \text{ and } f_{\bar{D}}(y)=f_Y(y)a_{\bar{D}}(y;x,\theta), \qquad \ldots(6.1)$$

where

$$a_D(y;x,\theta)=\Phi\left(\mu_{Y_D|Y}/\sigma_{Y_D|Y}\right)\Big/\Phi\left(\beta_{p+1}^T x_{P+1}\right),$$

$$a_{\bar{D}}(y;x,\theta)=\Phi(-\mu_{Y_D|Y}/\sigma_{Y_D|Y})/\Phi(-\beta_{p+1}^T x_{P+1}), \qquad \ldots(6.2)$$

$$\mu_{Y_D|Y}=\beta_{p+1}^T x_{p+1}+\sigma_{YY_D}^T\Sigma_{YY}^{-1}(y-(\beta_1^T x_1+\ldots+\beta_p^T x_p))$$

and

$$\sigma_{Y_D|Y}=(1-\sigma_{YY_D}^T\Sigma_{YY}^{-1}\sigma_{YY_D})^{1/2}.$$

Note that in our set up the likelihood ratio $\lambda(y)=f_D(y)/f_{\bar{D}}(y)$ is an increasing function of $s^*(y)=\sigma_{YY_D}^T\Sigma_{YY}^{-1}y$. Thus $s^*(y)$ is the optimum linear combination of the marker values in the sense of Theorem 2 above. Also, it is evident that the likelihood ratio $\lambda(y)$ depends on $P_D(D=1/y)/\{1-P_D(D=1/y)\}$. Our formulation thus leads to a discriminant type of function based on a kind of probit regression.

Now denoting the conditional distribution of $s^*(Y)$ given $D=1$ and $D=0$ by $f_D(s^*(y))$ and $f_{\bar{D}}(s^*(y))$ respectively we obtain

$$f_D(s^*(y)) = f_Y(s^*(y))a_D^*(y;x,\theta) \text{ and } f_{\bar{D}}(s^*(y)) = f_Y(s^*(y))a_{\bar{D}}^*(y;x,\theta),$$

where

$$a_D^*(y;x,\theta) = \Phi\left(\mu_{Y_D|s^*(y)} / \sigma_{Y_D|s^*(y)}\right) \Big/ \Phi\left(\beta_{p+1}^T x_{p+1}\right)$$

and

$$a_D^*(y;x,\theta) = \Phi\left(-\mu_{Y_D|s^*(y)} / \sigma_{Y_D|s^*(y)}\right) \Big/ \Phi\left(-\beta_{p+1}^T x_{p+1}\right).$$

6.2 Challenges in Estimation of the Confidence Intervals for Adjusted AUC

In Section 4.2, we have seen that the asymptotic variance $\sigma_H^2(x;\theta)$ of AUC is a function of $\bar{I}(\theta)$ and $H'_\theta(x;\theta)$, where the latter has a complicated expression and is not easy to obtain. As an alternative, we may estimate $\sigma_H^2(x;\theta)$ directly using paired bootstrap and then find normal theory bootstrap-based CI based on the facts that

(a) $n\hat{v}_H^{-1/2} / \sigma_H^2(x;\theta) \to 1 + O_{P_B}\left(n^{-1/2}\right)$, where $\hat{v}_H$ is the bootstrap estimate of $n^{-1}\sigma_H^2(x;\theta)$ and $P_B(.)$ represents bootstrap probability distribution given the sample data.

(b) $P\left[\hat{v}_H^{-1/2}\left(H\left(x;\hat{\theta}\right) - H(x;\theta)\right) \to N(0,1)\right] = 1.$

Now, for Monte-Carlo based Bootstrap estimation of $n^{-1}\sigma_H^2(x;\theta)$, we can, for example, consider B with replacement resamples of size n from the sampled data, where B is a large positive number. Based on B bootstrap samples we can calculate the variance estimate $\hat{v}_H$ as

$$\hat{v}_H \approx B^{-1}\sum_b\left(H\left(x;\hat{\theta}_b\right)-B^{-1}\sum_b H\left(x;\hat{\theta}_b\right)\right)^2,$$

which may now be used to obtain an appropriate confidence interval for adjusted AUC.

6.3 Performance when Sample Sizes are Small

We can see from the simulation exercise carried over in Section 5 that our proposed methodology performs better than the standard methodology. One question of interest is to see the effect of the sample size on these two competing methods. Looking at higher sample sizes is of course not very interesting as in that case both methods approach perfection and there is little to choose between the two methodologies. However, one interesting question is the relative performance of the two methods when the sample sizes are small, which is not too uncommon in medical studies.

To address this interesting question, we repeat our simulation exercise from Section 5, but this time with a sample size of 25. The average adjusted AUC values are presented in Table 2, and the ROC curves for the first, second and third quartiles of the covariate *x* are presented in Figures 3 and 4.

Table 2: The average adjusted AUC values for the two simulated data sets with sample size 25.

Values of the Covariate	Average Adjusted AUC values for n = 25			
	Simulation 1		Simulation 2	
	Standard Method	Proposed Method	Standard Method	Proposed Method
First Quartile	0.6805	0.7187	0.8914	0.9471
Second Quartile/ Median	0.6845	0.7044	0.8809	0.9449
Third Quartile	0.7157	0.7455	0.8691	0.9474

We note that the performance of the proposed method is consistently better for both the simulation exercises for all three quartiles. We would like to point out that the standard method was highly unstable with n = 25, especially for *x* = minimum or maximum when in many cases there were very few observations in one group. Hence we decided that it was not meaningful to compare the two methods in such cases.

We note that this observation is in sync with reality: extreme values of a meaningful covariate should indicate either the presence or the absence of the disease, and hence, especially when number of subjects in the sample is small, one of the groups may not be adequately represented in the sample. In that case the proposed methodology is actually much more stable than the standard method, and can still be employed meaningfully. Applicability when sample sizes are small is indeed one of the main advantages of the methodology we propose.

Figure 1: ROC *curves based on bivariate normal population of simulation* 1 *using proposed* (*blue line*) *and standard* (*red line*) *methodologies at extreme* x*-values* (*graph* 1 *for minimum and graph* 5 *for maximum*) *and three quartiles* (*graphs* 2-4 *for first -third quartiles*).

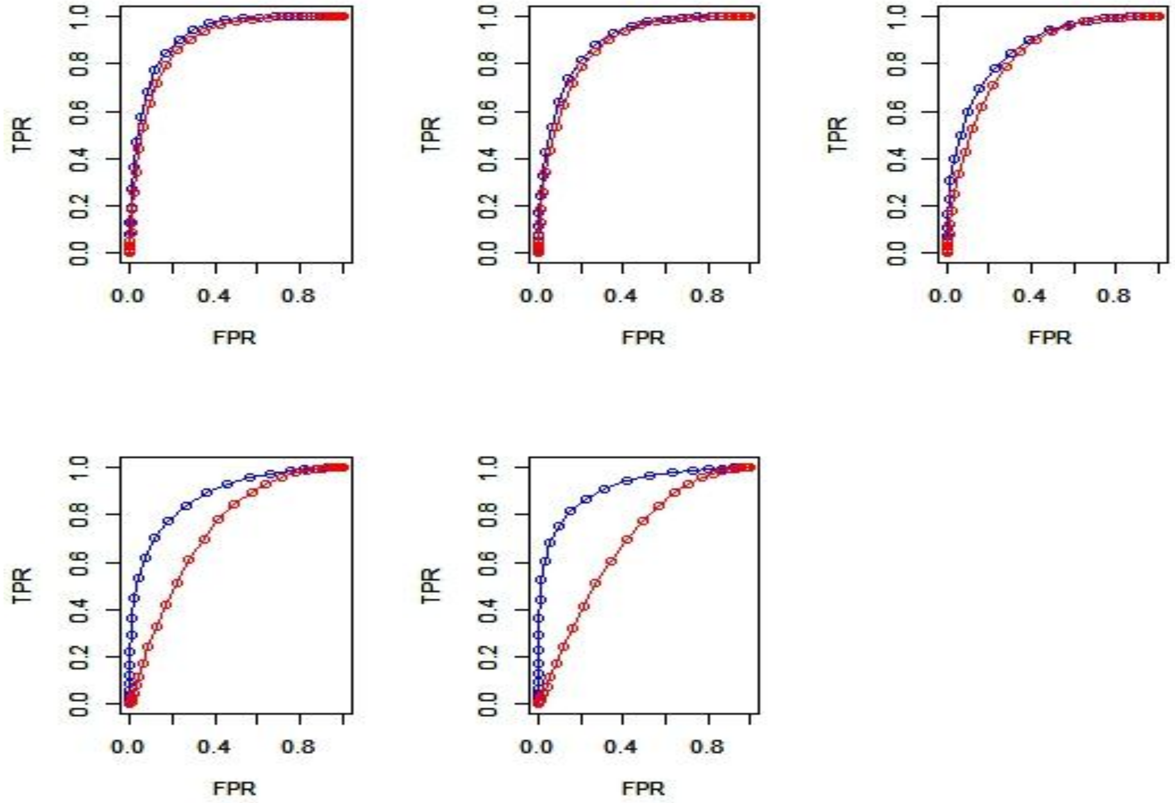


Figure 2: ROC *curves based on bivariate normal population of simulation* 2 *using proposed* (*blue line*) *and standard* (*red line*) *methodologies at extreme* x*-values* (*graph* 1 *for minimum and graph* 5 *for maximum*) *and three quartiles* (*graph* 2-4 *for first -third quartiles*)

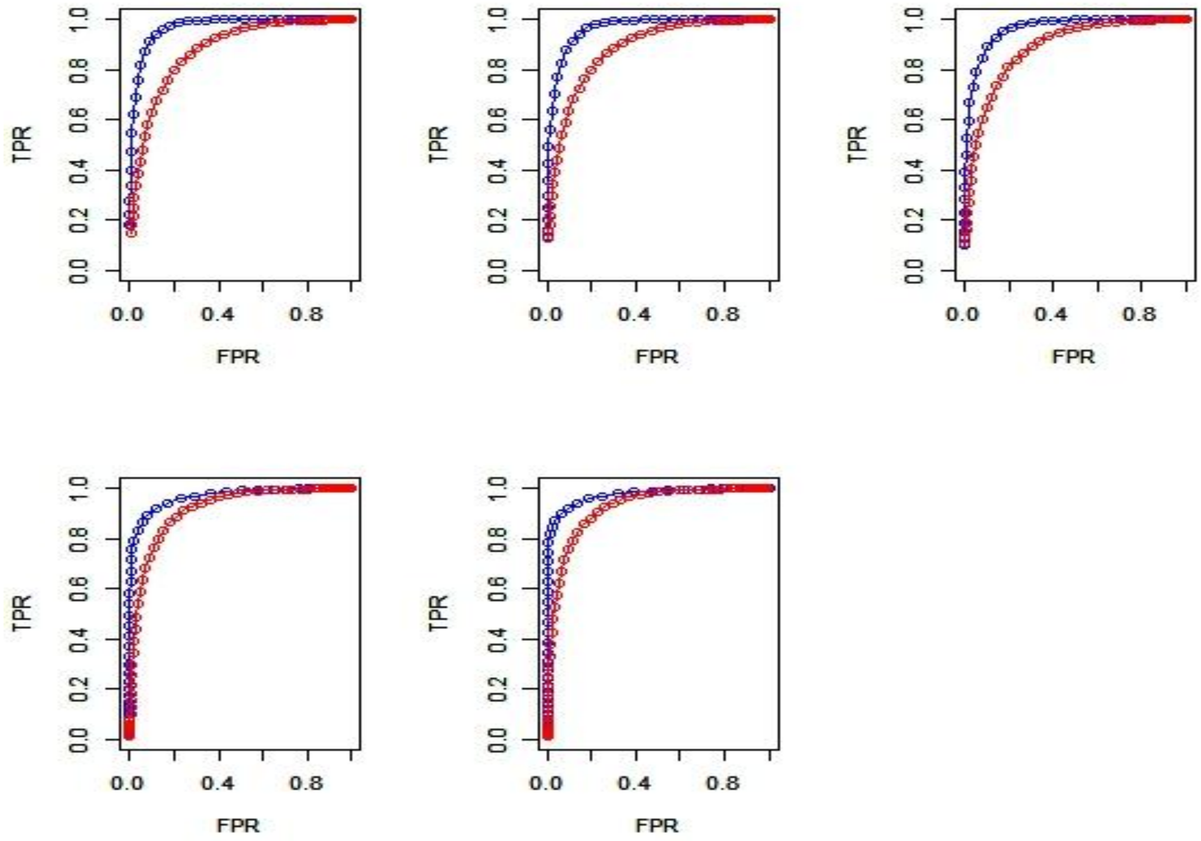


Figure 3: ROC *curves based on bivariate normal population of simulation* 1 *using proposed* (*blue line*) *and standard* (*red line*) *methodologies at three quartiles for sample size* 25.

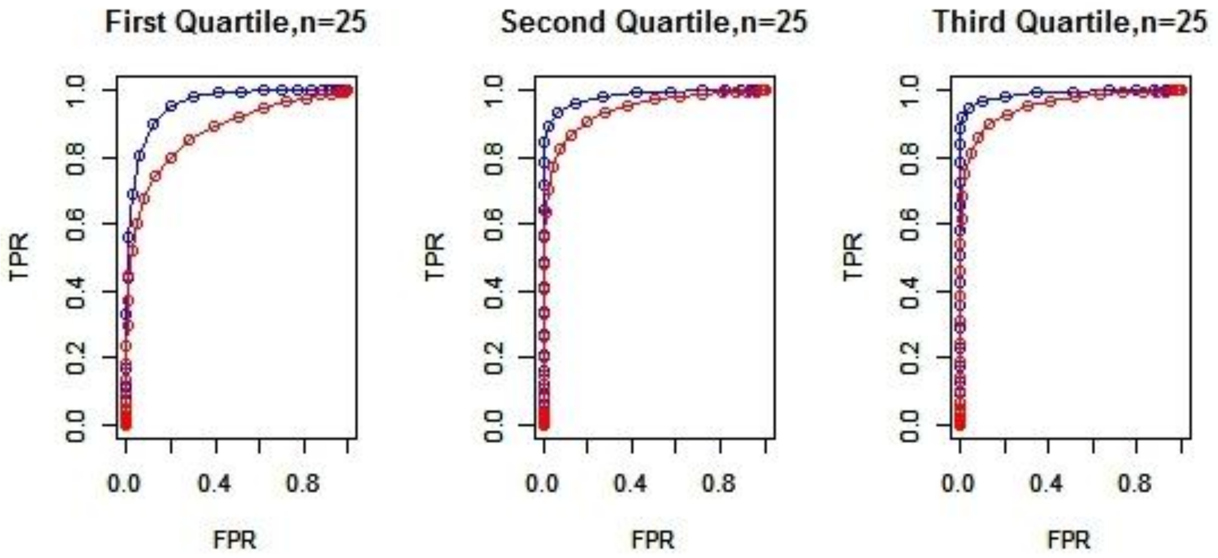
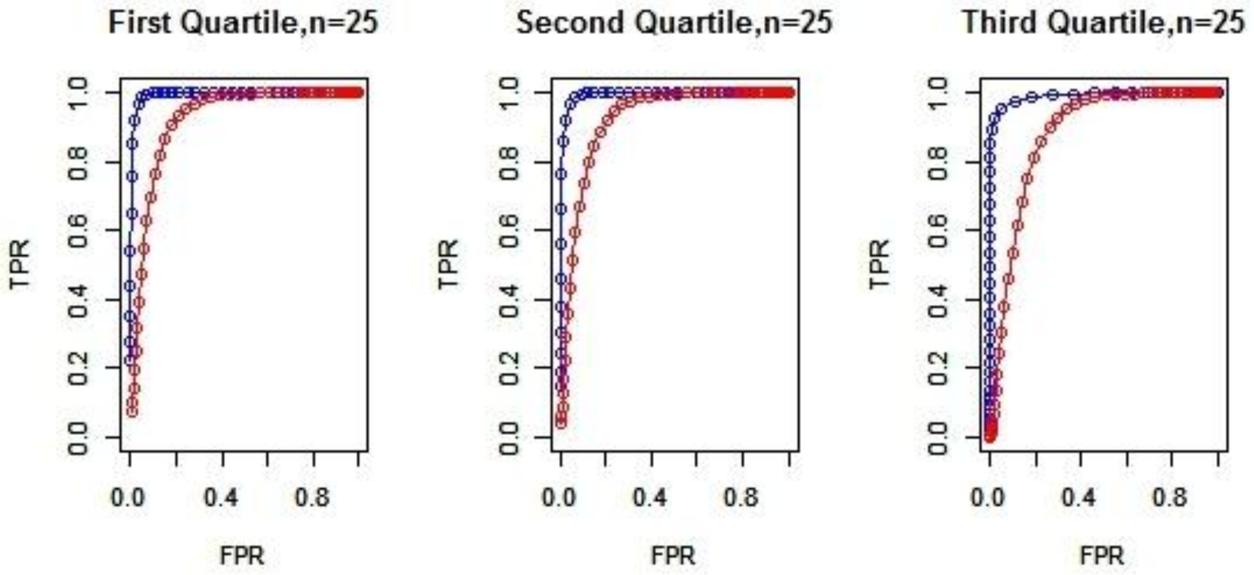


Figure 4: ROC *curves based on bivariate normal population of simulation* 2 *using proposed* (*blue line*) *and standard* (*red line*) *methodologies at three quartiles for sample size* 25.



References

Baker SG. The central role of receiver operating characteristic (ROC) curves in evaluating tests for the early detection of cancer. *Journal of the National Cancer Institute.* 2003; 95: 511‑515.

Bamber DC. The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*. 1975; 12:387‑415.

Chang YCI. Maximizing an ROC-type measure via linear combination of markers when the gold reference is continuous. *Statistics in medicine.* 2013; 32: 1893-1903.

Dorfman DD, Alf E. Maximum likelihood estimation of parameters of signal detection theory and determination of confidence intervals-rating method data. *Journal of Mathematical Psychology*. 1969; 6:487-96.

Janes H, Longton G and Pepe MS. Accommodating covariates in receiver operating characteristic analysis. *The Stata Journal*. 2009; 1: 17-39.

Janes H and Pepe MS. Matching in studies of classification accuracy: Implications for analysis, efficiency, and assessment of incremental value. *Biometrics*. 2008; 64:1–9.

Kerr KF and Pepe MS. Joint Modeling, Covariate Adjustment, and Interaction: Contrasting Notions in Risk Prediction Models and Risk Prediction Performance. *Epidemiology*. 2011; 22: 805–812.

Lehmann, EL. *Elements of large-sample theory*. Springer, New York, 1999.

Lin H, Zhou L, Peng H and Zhou XH. Selection and combination of biomarkers using ROC method for disease classification and prediction. *Canadian Journal of Statistics*. 2011; 39: 324-343.

Liu A, Schisterman EF and Zhu Y. On linear combinations of biomarkers to improve diagnostic accuracy. *Statistics in Medicine.* 2005; 24: 37–47.

Ma S and Hunag, J. Regularized ROC method for disease classification and biomarker selection with microarray data. *Bioinformatics.* 2005; 21: 4356–4362.

Ma J, Xiaofei W and Stephen G. Semiparametric estimation of ROC curve under test-result-dependent sampling. *Duke Biostatistics and Bioinformatics Working Paper Series.* 2010; Paper 9.

Metz C, Wang P and Kronman H. A new approach for testing the significance of differences between the ROC curves measured from correlated data. In: Deconick, F. (editor), *Information Processing in Medical Imaging VIII*. 1984; 432–445.

Ogilvie JC and Creelman CD. Maximum likelihood estimation of ROC curve parameters. *Journal of Mathematical Psychology*. 1968; 5:377-91.

Pardo-Fernández JC, Rodrıguez-Alvarez MX and Van Keilegom I. (2014). A review on ROC curves in the presence of covariates. *REVSTAT–Statistical Journal*. 2014; 12, 21-41.

Pepe MS. *The statistical evaluation of medical tests for classification and prediction*. Oxford University Press, 2003.

Pepe MS, Fan J, Seymour CW, Li C, Huang Y and Feng Z. Biases Introduced by Choosing Controls to Match Risk Factors of Cases in Biomarker Research. *Clinical Chemistry*. 2012; 58: 1242–1251.

Pepe MS, Etzioni R, Feng Z, Potter JD, Thompson ML, Thornquist M, Winget M and Yasui Y. Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institut*e. 2001; 93: 1054‑1061.

Schisterman EF, David Faraggi D and Reiser B. Adjusting the generalized ROC curve for covariates. *Statistics in Medicine*. 2004; 23:3319‑3331.

Wang Z and Chang YCI. Marker selection via maximizing the partial area under the ROC curve of linear risk scores. *Biostatistics*. 2011; 12: 369-385.

Wang Z, Chang Y, Ying Z, Zhu L and Yang Y. A parsimonious threshold-independent protein feature selection method through the area under receiver operating characteristic curve. *Bioinformatics*. 2007; 23: 2788–2794.

Zhou C, Obuchowski N and McClish D. *Statistical Methods in Diagnostic Medicine*. New York: Wiley, 2002.