University of Michigan School of Public Health
The University of Michigan Department of Biostatistics Working Paper Series

Year 2004 *Paper* 47

Semiparametric methods for the binormal model with multiple biomarkers

Debashis Ghosh*

*University of Michigan, debashis.ghosh@ucdenver.edu

http://biostats.bepress.com/umichbiostat/paper47

Semiparametric methods for the binormal model with multiple biomarkers

Debashis Ghosh

Abstract

Abstract: In diagnostic medicine, there is great interest in developing strategies for combining biomarkers in order to optimize classification accuracy. A popular model that has been used when one biomarker is available is the binormal model. Extension of the model to accommodate multiple biomarkers has not been considered in this literature. Here, we consider a multivariate binormal framework for combining biomarkers using copula functions that leads to a natural multivariate extension of the binormal model. Estimation in this model will be done using rank-based procedures. We also discuss adjustment for covariates in this class of models and provide a simple two-stage estimation procedure that can be fit using standard software packages. Some analytical comparisons between analyses using the proposed model with univariate biomarker analyses are given. In addition, the techniques are applied to simulated data as well as data from two cancer biomarker studies.

Semiparametric methods for the binormal model with multiple biomarkers

Debashis Ghosh

Department of Biostatistics, University of Michigan

1420 Washington Heights

Ann Arbor, MI 48109-2029

Abstract

In diagnostic medicine, there is great interest in developing strategies for combining biomarkers in order to optimize classification accuracy. A popular model that has been used when one biomarker is available is the binormal model. Extension of the model to accommodate multiple biomarkers has not been considered in this literature. Here, we consider a multivariate binormal framework for combining biomarkers using copula functions that leads to a natural multivariate extension of the binormal model. Estimation in this model will be done using rank-based procedures. We also discuss adjustment for covariates in this class of models and provide a simple two-stage estimation procedure that can be fit using standard software packages. Some analytical comparisons between analyses using the proposed model with univariate biomarker analyses are given. In addition, the techniques are applied to simulated data as well as data from two cancer biomarker studies.

Keywords: Dependence, Linear Regression, Multivariate distribution, Screening, Transformation Model.

1. Introduction

In most medical settings, it is becoming increasingly clear that one biomarker will not be sufficient to serve as a screening device for early detection of many diseases. As an example, we consider prostate cancer. Typically, prostate-specific antigen (PSA) has been used for detection of the disease. If a man has a PSA measurement between 4 and 10 ng/mL, then this leads to a prostate needle biopsy. While PSA is known for being a relatively sensitive biomarker, it is not known as being a very specific measurement. As a result, many biopsies yield negative results for tumor, even when the PSA is between 4-10 ng/mL. A current estimate of the specificity of PSA is approximately 35% (Kawinski et al., 2002). Many investigators believe that a combination of biomarkers will potentially lead to more sensitive screening rules for detecting prostate cancer. How best to combine these measurements remains an open question.

There has been much recent work in terms of developing methods for combining multiple biomarkers. Su and Liu (1993) and Pepe and Thompson (2000) considered linear combinations of biomarkers to optimize measures of diagnostic accuracy. McIntosh and Pepe (2002) noted the optimality of the likelihood ratio, and by Bayes Theorem, the risk score, in terms of developing sensitive screening rules. We outline some of their results in Section 2. In addition, Baker (2000) developed an algorithmic approach for finding combinations of biomarkers using the likelihood ratio. His arguments were motivated more by results from decision theory and cost-effectiveness analysis (Weinstein et al., 1980) than the Neyman-Pearson approach of McIntosh and Pepe (2002). Recently, Etzioni et al. (2003) proposed developing screening rules based on consideration of logical combinations of biomarker measurements.

In the case of one biomarker, there has been extensive work done on the development of methodology for diagnostic testing and screening (Zhou et al., 2002; Pepe, 2003). An important quantity in this area is the receiver operating characteristic (ROC) curve, which is a plot of the true positive rate versus the false positive rate. A popular approach to modelling the ROC curve for one biomarker is the binormal model (Swets, 1986; Pepe 2003, §4.4). The model plays a central role in modelling of ROC curves. However, there currently exist no multivariate versions of the binormal model. A multivariate binormal model, if available,

would allow for an alternative method of combining biomarkers relative to the approaches listed in the previous paragraph.

In this article, we develop a multivariate extension of the binormal model using copula functions (Nelsen, 1999). While copulas have chiefly played a role in the analysis of correlated survival data (e.g., Oakes, 1989), it turns out that they can be applied here as well. The structure of this paper is as follows. In Section 2, we provide some background on ROC curves, the binormal model and copula functions. We formulate the multivariate binormal model in this framework and discuss estimation procedures for this model in Section 3. There, we also consider strategies for covariate adjustment in this model and propose a new two-stage estimation procedure for the multivariate binormal model. We also develop asymptotic results of the proposed methods. In Section 4, we examine the potential gains in terms of a multivariate analysis relative to a univariate biomarker analysis. There, the proposed methodology is applied to simulated data as well as data from two cancer biomarker studies. We conclude with some brief remarks in Section 5.

2. Data and Background

We will be assuming that we have data $(D_i, \mathbf{Y}_i, \mathbf{X}_i)$, $i = 1, \ldots, n$, a random sample from $(D, \mathbf{Y}, \mathbf{X})$, where D denotes the disease status, $\mathbf{Y} \equiv (Y_1, \ldots, Y_p)$ is a p-dimensional biomarker measurement, and $\mathbf{X}$ is a q-dimensional vector of covariates.

2.1 ROC curve and binormal curve for one biomarker

Suppose we have only biomarker Y. We will assume that higher values of the biomarker correspond to a greater probability of having disease. One relevant quantity is the false positive rate based on a cutoff c, defined to be $FP(c) = P(Y > c|D = 0)$. Similarly, the true positive rate is $TP(c) = P(Y > c|D = 1)$. The true and false positive rates can be summarized by the receiver operating characteristic (ROC) curve, which is a graphical presentation of $\{TP(c), FP(c) : -\infty < c < \infty\}$. The ROC curve shows the tradeoff between increasing true positive and false positive rates. Tests that are have $\{TP(c), FP(c)\}$ values close to (0,1) indicate perfect discriminators, while those with $\{TP(c), FP(c)\}$ values close

to the 45^o degree line in the $(0,1) \times (0,1)$ plane are tests that are unable to discriminate between the diseased and healthy populations.

Suppose that the biomarker distribution in the diseased population is assumed to be normal with mean μ_D and variance σ_D^2, while that in the undiseased population is normal with mean μ_U and variance σ_U^2. Then in Pepe (2003, p. 82), it is shown that for $t \in [0,1]$,

$$\text{ROC}(t) = \Phi\{a + b\Phi^{-1}(t)\}, \tag{1}$$

where $a = (\mu_D - \mu_U)/\sigma_D$ and $b = \sigma_U/\sigma_D$. Equation (1) is referred to as the binormal ROC curve. Several authors (Swets, 1986; Hanley, 1988) have demonstrated that the binormal ROC curve provides a good approximation to many empirical ROC curves that occur in practice. As described in Pepe (2003, p. 81), "the binormal ROC curve plays a central role in ROC analysis. In much the same way that the normal distribution is a classic model for distribution functions, the binormal ROC curve is a classic model for ROC curves."

It turns out that if h is a monotone transformation, then the ROC curve for the induced biomarker measurement $h(Y)$ is still (1). Thus, the binormal ROC curve model can be equivalently formulated as there existing a monotone transformation h such that $h(Y)$ has normal distributions with mean μ_D and μ_U and variance σ_D^2 and σ_U^2 in the diseased and undiseased populations, respectively. Because the monotone transformation h is identifiable only up to location and scale, we will assume without loss of generality that $\mu_D = 0$ and $\sigma_D^2 = 1$. It is this model that we generalize in §3.

2.2 Combining biomarkers

Suppose now we have more than one biomarker; we seek to combine information from multiple biomarkers in order to discriminate between diseased and healthy populations. A justification for combining biomarker measurements using the likelihood ratio from the ROC point of view was recently put forward by McIntosh and Pepe (2002). They provide a reinterpretation of the Neyman-Pearson lemma and show that the classification rule based on

$$\text{LR}(\mathbf{Y}) \equiv \frac{Pr(\mathbf{Y}|D=1)}{Pr(\mathbf{Y}|D=0)} > c(f_0) \tag{2}$$

optimizes the sensitivity for a given false positive rate f_0, where $c(f_0)$ is chosen such that $Pr\{LR(\mathbf{Y}) > c(f_0)\} = f_0$, and $f_0 \in [0,1]$. McIntosh and Pepe (2002) refer to this optimality property as the uniformly most sensitive (UMS) sensitive screening test based on $\mathbf{Y} \equiv (Y_1, \ldots, Y_p)$. They also mention two other optimality properties of screening tests based on $LR(\mathbf{Y})$. First, tests of the form (2), for a given f_0, minimize the overall misclassification rate. Second, tests of the form (2) minimize expected cost, where unequal costs are given to false positives and false negatives.

By Bayes' rule, McIntosh and Pepe (2002) show that

$$P(D = 1|\mathbf{Y}) \quad = \quad \frac{LR(\mathbf{Y})\pi}{LR(\mathbf{Y})\pi + 1},$$

where $\pi \equiv P(D = 1)/P(D = 0)$ is the odds of disease in the population. This implies that (2) can be written as $P(D = 1|\mathbf{Y}) > \tilde{c}(f_0)$ for $\tilde{c}(f_0) = c(f_0)\pi/\{c(f_0)\pi + 1\}$. Thus, as McIntosh and Pepe (2002) argue, one can construct UMS rules based on the risk score $P(D = 1|\mathbf{Y})$. If one assumes that $P(D = 1|\mathbf{Y})$ is of the form

$$\frac{\exp(\beta_0 + \beta^T\mathbf{Y})}{1 + \exp(\beta_0 + \beta^T\mathbf{Y})},$$

then logistic regression can be used to estimate UMS rules. Another advantage is that the regression models can be fit to case-control data in order to derive optimal screening rules.

There are two modelling issues that arise from this framework. If we attempt to develop rules based on modelling $P(\mathbf{Y}|D)$, then this leads to specification of complex multivariate distributions; incorporation of covariates is even more complex. If we instead choose to model the risk score, $P(D = 1|\mathbf{Y})$, then consideration of interactions and more generally, model selection issues arise here. However, incorporating covariates is straightforward, as one can specify a model for $P(D = 1|\mathbf{Y}, \mathbf{X})$.

Copula functions offer advantages relative to these two approaches. In particular, these functions allow for flexible modelling of univariate biomarkers and covariate adjustment on disease outcome, which is a well-characterized area. The dependence between biomarkers is specified parametrically in copula functions. In most situations, there is sufficient data available to characterize marginal distributions of variables but less information on estimating interrelationships between them. Copula models are attractive in that nonparametric

modelling procedures can be considered for the marginal distribution while the interactions between biomarkers are modelled in a parametric fashion. Thus, copula models are a natural way of modelling multivariate data. Before describing the extension of the binormal model using copula functions, we provide some background on them.

2.3 Copula functions

For the sake of exposition, we assume that there are no covariates and only two available biomarkers, Y_1 and Y_2. A copula model links the joint distribution of Y_1 and Y_2, conditional on disease status, to the marginal conditional distributions:

$$Pr(Y_1, Y_2|D) = C_\theta\{Pr(Y_1|D), Pr(Y_2|D)\}, \tag{3}$$

where C_θ is a function that maps from $[0,1] \times [0,1]$ to $[0,1]$, and θ is a dependence parameter. Copula functions have been utilized with success in the analysis of censored survival data (Oakes, 1989, Hsu and Prentice, 1995). The most popular copula model in that area is based on the Clayton-Oakes frailty model (Clayton, 1978; Oakes, 1986):

$$Pr(Y_1, Y_2|D) = \{Pr(Y_1|D)^{-\theta} + Pr(Y_2|D)^{-\theta} - 1\}^{-1/\theta}. \tag{4}$$

In (4), θ is a dependence parameter that takes values $[-1, \infty)$, although the joint distribution of Y_1 and Y_2 is absolutely continuous for $\theta > -1/2$. This parameter has an interpretation as a cross-ratio function (Oakes, 1989). Other choices of copula functions are available; a comprehensive summary of such models is available in §4.1 of Nelsen (1999).

Although the copula specification may appear unfamiliar at first glance, there are in fact many methods of analysis that either implicitly or explicitly involve copula functions or related quantities (e.g., Dale, 1986; Molenberghs and Lesaffre, 1994; Heagerty and Zeger, 1996).

3. Proposed Methodology

3.1 Model and estimation: no covariates

We now extend the binormal model from §2.1 to accommodate multiple markers. Generalizing the model discussed there, we assume that there exist monotone transformations $G_1, \ldots, G_p$ such that the vector $\mathbf{G}(\mathbf{Y}) \equiv \{G_1(Y_1), \ldots, G_p(Y_p)\}$ has a multivariate normal distribution with mean $\mu_D \equiv (\mu_{D1}, \ldots, \mu_{Dp})$ and variance-covariance matrix $\mathbf{\Sigma}_D$ among the diseased population and mean $\mu_{\bar{D}} \equiv (\mu_{\bar{D}1}, \ldots, \mu_{\bar{D}p})$ and variance-covariance matrix $\mathbf{\Sigma}_{\bar{D}}$ among the nondiseased population. Thus, after monotone transformations performed marginally on the Y values, multivariate normality is assumed in both populations. This model was studied by Lin and Jeon (2003) as a semiparametric competitor to classical linear and quadratic discriminant analysis methods. We are providing an alternative motivation of this model using ROC ideas. In addition, we will cast this model into the copula framework described in §2.3 and will discuss covariate adjustment in this model in §3.2. Note that the monotone transformations will be unique up to scale and shift, so there will be no loss of generality in taking the marginal distributions for the biomarker measurements for the diseased population to have zero mean and variance one. Before discussing estimation, we cast the multivariate binormal model into the copula framework.

Let F denote the joint distribution of $\mathbf{Y}$. We assume that there exist monotone transformations $G_1, \ldots, G_p$ such that

$$F(\mathbf{y}|D) = C_{\theta,D}[\Phi\{G_1(y_1)\}, \ldots, \Phi\{G_p(y_p)\}] \tag{5}$$

and

$$F(\mathbf{y}|\bar{D}) = C_{\theta,\bar{D}}\left[\Phi\left\{\frac{G_1(y_1) - \mu_{\bar{D}1}}{\sigma_{\bar{D}1}}\right\}, \ldots, \Phi\left\{\frac{G_p(y_p) - \mu_{\bar{D}p}}{\sigma_{\bar{D}p}}\right\}\right], \tag{6}$$

where $\Phi(z)$ represents the cumulative distribution function for a standard normal random variable and C_θ is a multivariate extension of the 2-dimensional copula function described in §2.3. The extension of the univariate binormal model mentioned in §2.1 involves use of the p-dimensional normal or Gaussian copula function $C_\Phi(\mathbf{u})$; for $\mathbf{u} \in [0,1]^p$, the joint density corresponding to the copula function is given by

$$c_\Phi(\mathbf{u}|\mathbf{R}) = |\mathbf{R}|^{-1/2} \exp\left\{-\frac{1}{2}\mathbf{s}^T(\mathbf{R}^{-1} - \mathbf{I})\mathbf{s}\right\}, \tag{7}$$

where $\mathbf{s} \equiv \{\Phi^{-1}(u_1), \ldots, \Phi^{-1}(u_p)\}^T$ and Φ^{-1} is the inverse function of Φ. The $\mathbf{R}$ is a cor-

relation matrix where the off-diagonal elements specify the dependence between biomarkers. The model is semiparametric because $G_1, \ldots, G_p$ are unspecified infinite-dimensional nuisance parameters while the components of $\mu_{\bar{D}}$, $\sigma^2_{\bar{D}} \equiv (\sigma^2_{\bar{D}1}, \sigma^2_{\bar{D}2}, \ldots, \sigma^2_{\bar{D}p})$ and $\mathbf{R}$ are the finite dimensional parameters. Note that there will be a separate $\mathbf{R}$, $\mathbf{R}_D$ and $\mathbf{R}_{\bar{D}}$, for the diseased and nondiseased populations, respectively. Estimation of the models (5) and (6) requires estimation of $\mu_{\bar{D}}$, the marginal variances $\sigma^2_{\bar{D}}$ and the correlation matrices $\mathbf{R}_D$ and $\mathbf{R}_{\bar{D}}$.

Suppose that $G_1, \ldots, G_p$ were known. Then if Y_j denotes the jth biomarker ($j = 1, \ldots, p$), $G_j(Y_j)$ is marginally distributed as a $N(0,1)$ random variable in the diseased population, $j = 1, \ldots, p$. We can then estimate G_j ($j = 1, \ldots, p$) as $\hat{G}_j = \Phi^{-1} \circ \hat{F}^D_j$, where $h \circ f$ is the composition of two functions h and f and $\hat{F}^D_j$ is the empirical cumulative distribution of Y_j in the diseased population, $j = 1, \ldots, p$. Based on this result, we can then develop the following simple method of moments estimator of $\mu_{\bar{D}j}$ ($j = 1, \ldots, p$) following arguments similar to those in Lin and Jeon (2003):

$$\hat{\mu}_{\bar{D}j} = \frac{\sum_{i=1}^n \hat{G}_j(Y_{ij}) I(D_i = 0)}{\sum_{i=1}^n I(D_i = 0)},$$

where Y_{ij} is the jth component of $\mathbf{Y}_i$, $i = 1, \ldots, n$. Similarly, the marginal variance for the jth biomarker ($j = 1, \ldots, p$), $\sigma^2_{\bar{D}j}$, can be estimated using method of moments:

$$\hat{\sigma}^2_{\bar{D}j} = \frac{\sum_{i=1}^n (\hat{G}_j(Y_{ij}) - \hat{\mu}_{\bar{D}j})^2 I(D_i = 0)}{\sum_{i=1}^n I(D_i = 0) - 1}.$$

In practice, we will use trimmed estimators of these quantities so that they are more robust and do not depend heavily on the amount of overlap in the biomarker distributions between diseased and undiseased individuals. We follow the recommendations of Lin and Jeon (2003) for trimming.

It now remains to estimate the correlation matrices. For $i, j = 1, \ldots, p$, the estimate of the (j,k)th element of $\mathbf{R}_D$, ρ^D_{jk}, is given by

$$\hat{\rho}^D_{jk} = \frac{\sum_{i=1}^n \hat{G}_k(Y_{ik}) \hat{G}_j(Y_{jk}) I(D_i = 1)}{[\sum_{i=1}^n \{\hat{G}_k(Y_{ik}) I(D_i = 1)\}^2]^{1/2} [\sum_{i=1}^n \{\hat{G}_j(Y_{ij}) I(D_i = 1)\}^2]^{1/2}}. \quad (8)$$

Similarly, the estimator of the (j,k)th element of $\mathbf{R}_{\bar{D}}$ is given by

$$\hat{\rho}^{\bar{D}}_{jk} = \frac{\sum_{i=1}^n \{\hat{G}_k(Y_{ik}) - \hat{\mu}_{\bar{D}k}\}\{\hat{G}_j(Y_{ij}) - \hat{\mu}_{\bar{D}j}\} I(D_i = 0)}{\sum_{i=1}^n [\{\hat{G}_k(Y_{ik}) - \hat{\mu}_{\bar{D}k}\}^2 I(D_i = 0)^2]^{1/2} [\sum_{i=1}^n \{\hat{G}_j(Y_{ij}) - \hat{\mu}_{\bar{D}j}\}^2 I(D_i = 0)^2]^{1/2}}, \quad (9)$$

The formulae (8) and (9) is referred to as the Van der Waerden normal scores rank correlation coefficient (Klaassen and Wellner, 1997). In the case of $p = 1$ for one population, they show that $n^{1/2}(\hat{\rho} - \rho)$ has a limiting normal distribution with mean zero and variance $(1 - \rho)^2$, which is the semiparametric efficiency bound.

3.2 Model and estimation: covariates

In practice, the distributions of the biomarkers in the diseased and nondiseased populations will depend on other covariates. For example, if a longitudinal series of biomarkers is used for screening for disease, then one might need to adjust for time of measurement for diseased populations while not adjust for it in undiseased populations. In addition, if severity of disease is associated with the discriminative ability of the biomarker, then we will want to adjust for stage of disease in the diseased subjects but not in the undiseased populations. Thus, we will want to consider multivariate binormal models in which we adjust for covariates.

We model the effect of covariates on the biomarkers can be using Box-Cox regression models; for the jth biomarker ($j = 1, \ldots, p$), for $D_i = 1$,

$$\frac{Y_{ij}^{\lambda_j} - 1}{\lambda_j} = \mathbf{X}_i^T \beta_{Dj} + \epsilon_{Dij}, \tag{10}$$

where $i = 1, \ldots, n$, β_{Dj} is a vector of unknown regression coefficients for the jth biomarker, $\lambda_j \in (0, \infty)$ and ϵ_{Dij} are iid error terms with mean zero, given j. However, given $i = 1, \ldots, n$, $(\epsilon_{Di1}, \ldots, \epsilon_{Dip})$ come from the normal copula model in §3.1. An analogous model exists for the undiseased population:

$$\frac{Y_{ij}^{\lambda_j} - 1}{\lambda_j} = \mathbf{X}_i^T \beta_{\bar{D}j} + \epsilon_{\bar{D}ij}, \tag{11}$$

where $\beta_{\bar{D}j}$ is a vector of unknown regression coefficients for the jth biomarker, $\lambda_j \in (0, \infty)$ and $\epsilon_{\bar{D}ij}$ are iid error terms with mean zero, given j. Similarly, given i, the distribution of $(\epsilon_{\bar{D}i1}, \ldots, \epsilon_{\bar{D}ip})$ is given by a normal copula model.

Note that fitting models (10) and (11) for each of the biomarkers allows for flexible incorporation of covariates. Although we have included the same covariates in the two models for each biomarker, this formulation can allow for different covariates for each individual

biomarker. In addition, we can include different covariates for the diseased and nondiseased subject populations. We will assume that λ is treated as fixed, although we can generalize to allow for a separate $(\lambda_1, \ldots, \lambda_p)$ for the diseased and nondiseased populations. By fitting separate models, we are also implicitly making the assumption that disease status interacts with $\mathbf{X}$. It should be noted that (10) and (11) are semiparametric models because other than having mean zero, the marginal distributions of the error terms remain unspecified.

We now describe our estimation approach. For purposes of exposition, we focus on the diseased population, the case for the nondiseased would follow similarly. First, we estimate the regression coefficients in (10) and (11) using least squares. Next, the residuals

$$r_{Dij} \equiv \frac{Y_{ij}^{\lambda} - 1}{\lambda} - \mathbf{X}_j^T \hat{\beta}_{Di}$$

and

$$r_{\bar{D}ij} \equiv \frac{Y_{ij}^{\lambda} - 1}{\lambda} - \mathbf{X}_j^T \hat{\beta}_{\bar{D}i}$$

are utilized for the estimation procedure described in Section 3.1. It turns out that the residuals in the diseased and undiseased populations do not involve the nonparametric component of the model. We prove this fact in the Appendix. In principle, this allows for easy development of the asymptotic properties of the estimators using the theory of U-statistics (Van der Vaart, 2000). Note that with this two-stage procedure, we seek to determine if there is any discriminatory power of the biomarker after adjusting for covariates.

3.3 Relationship with ROC methodology

An implication of the copula model (7) is that it preserves the marginal structure of the binormal model in §2.1. Thus, for the jth biomarker ($j = 1, \ldots, p$), the corresponding marginal ROC curve is binormal, i.e. for $t \in [0, 1]$,

$$\mathrm{ROC}_j(t) = \Phi\{a_j + b_j \Phi^{-1}(t)\}, \tag{12}$$

where $a_j = -\mu_{\bar{D}j}$ and $b_j = \sigma_{\bar{D}j}$, for the jth biomarker, $j = 1, \ldots, p$. Note that the ROC curve for the jth biomarker, summarized in (12) is of the same form as the ROC curve given by equation (1), where $\mu_{Dj} = 0$ and $\sigma_{Dj} = 1$ by assumption here. This shows how the

multivariate normal copula model presented here generalizes the univariate binormal model given in §2.1.

It is fruitful to consider the relationship of the proposed methodology with other methods for modelling the ROC curve. Let us consider $p = 1$ biomarker. Suppose we have the same covariates in the models (11) and (10) for the diseased and nondiseased populations, respectively. Recall the definition of the ROC curve, conditional on covariates (Pepe, 2000):

$$ROC_{\mathbf{Z}_U,\mathbf{Z}_D}(t) = P(Y_D > Y_U | F_U(Y_U) = t, \mathbf{Z}_U, \mathbf{Z}_D), \tag{13}$$

where Y_D is the biomarker measurement for a diseased individual, Y_U is the biomarker measurement for an undiseased individual, F_U is the cumulative distribution function for biomarker measurements in the undiseased population, and $\mathbf{Z}_D$ and $\mathbf{Z}_U$ are the covariate vectors for diseased and undiseased individuals. Then by algebraic manipulations, formulation of models (11) and (10) implies the following regression model for (13):

$$ROC_{\mathbf{Z}_U,\mathbf{Z}_D}(t) = \Phi\{\Phi^{-1}(t) + (\mathbf{Z}_D - \mathbf{Z}_U)^T\beta\}, \tag{14}$$

which is the class of models considered by Pepe (2000) in the case of one biomarker. Extending this argument, we have that for the multivariate version of the binormal model presented in §3.2, the ROC curve for the jth biomarker ($j = 1, \ldots, p$) will be of the form (14). Thus, the model presented here presents a method of extending the work of Pepe (2000) to accommodate multiple biomarkers.

Note that model (14) is more parsimonious in that effects of covariates are specified directly on the ROC curve. An alternative approach to modelling multivariate biomarkers is to specify the effects of covariates on the marginal ROC curve for each biomarker separately. We address this issue further in Section 5.

3.4 Combining biomarkers, classification and model evaluation

One way of conceptualizing the method proposed here is that after estimating the transformations $G_1, \ldots, G_p$, we perform a discriminant analysis in order to build a classification rule. The estimated rule is based on the linear discriminant function

$$LDF(\mathbf{Y}) = -\frac{1}{2}\hat{G}(\mathbf{Y})^T\hat{\mathbf{R}}_D\hat{G}(\mathbf{Y}) + \frac{1}{2}\{\hat{G}(\mathbf{Y}) - \hat{\mu}_{\bar{D}}\}^T\hat{\mathbf{S}}_{\bar{D}}\hat{\mathbf{R}}_{\bar{D}}\{\hat{G}(\mathbf{Y}) - \hat{\mu}_{\bar{D}}\}, \tag{15}$$

where $\hat{\mathbf{S}}_{\bar{D}}$ is a $p \times p$ diagonal matrix with $\hat{\sigma}^2_{\bar{D}}$ on the diagonal. The linear discriminant function is the method by which biomarkers are combined in the multivariate binormal model. This then provides a rule for classifying a subject as diseased or not diseased based on a multivariate biomarker profile. To assess the discriminatory power of the rule, it is useful to construct a ROC curve. Based on all possible cutpoints c^* for LDF, we can calculate the following two quantities:

$$\widehat{TPR}(c^*) = \frac{\sum_{i=1}^n I\{LDF(\mathbf{Y}_i) > c^*, D_i = 1\}}{\sum_{i=1}^n I(D_i = 1)}$$

and

$$\widehat{FPR}(c^*) = \frac{\sum_{i=1}^n I\{LDF(\mathbf{Y}_i) > c^*, D_i = 0\}}{\sum_{i=1}^n I(D_i = 0)}.$$

A plot of $\{\widehat{TPR}(c^*), \widehat{FPR}(c^*)\}$ then provides an ROC curve for the classification rule from the estimated multivariate binormal model.

Note that the ROC curve will be overoptimistic in that the predictions for $\widehat{TPR}$ and $\widehat{FPR}$ were based on the estimated multivariate binormal model, the parameter estimates of which were computed using the entire dataset. To reduce this overfitting, an alternative is to construct a leave-one-out cross-validation estimate of the ROC curve. Let $\widehat{LDF}^{(i)}(\mathbf{Y}_i)$ denote the estimate of LDF with the ith observation held out, $i = 1, \ldots, n$. Then a cross-validated estimate of the ROC curve is given by $\{\widetilde{TPR}(c^*), \widetilde{FPR}(c^*)\}$, where

$$\widetilde{TPR}(c^*) = \frac{\sum_{i=1}^n I\{\widehat{LDF}^{(i)}(\mathbf{Y}_i) > c^*, D_i = 1\}}{\sum_{i=1}^n I(D_i = 1)}$$

and

$$\widetilde{FPR}(c^*) = \frac{\sum_{i=1}^n I\{\widehat{LDF}^{(i)}(\mathbf{Y}_i) > c^*, D_i = 0\}}{\sum_{i=1}^n I(D_i = 0)}.$$

4. Numerical Comparisons

4.1. Analytical Results and Simulation Studies

In this section, we consider a comparison of the multivariate binormal normal relative to a univariate biomarker analysis. First, an analytical comparison is performed. We consider the case of $p = 2$ biomarkers and no covariates and where G_1 and G_2 are the identity functions.

This leads to a simple bivariate normal model in which $\mathbf{Y}|D = 1$ is distributed bivariate normal with mean zero vector and correlation matrix

$$\mathbf{\Sigma}_D \equiv \begin{pmatrix} 1 & \rho_D \\ \rho_D & 1 \end{pmatrix}$$

while $\mathbf{Y}|D = 0$ is distributed bivariate normal with mean vector $\mu_{\bar{D}} \equiv (\mu_{\bar{D}1}, \mu_{\bar{D}2})$ and covariance matrix

$$\mathbf{\Sigma}_{\bar{D}} \equiv \begin{pmatrix} \sigma^2_{\bar{D}1} & \sigma_{\bar{D}1}\sigma_{\bar{D}2}\rho_{\bar{D}} \\ \sigma_{\bar{D}1}\sigma_{\bar{D}2}\rho_{\bar{D}} & \sigma^2_{\bar{D}2}. \end{pmatrix}$$

Suppose we use the area under the ROC curve for assessing discriminatory power for comparing an analysis based on the first biomarker versus the multivariate binormal model. Some straightforward algebra yields that the area under the ROC curve for the univariate analysis is

$$AUC_1 = \Phi\left\{\left(\frac{\mu^2_{\bar{D}1}}{1+\sigma^2_{\bar{D}1}}\right)^{1/2}\right\}$$

while the corresponding quantity for the multivariate binormal model is

$$AUC_M = \Phi\left[\left\{\mu^T_{\bar{D}}(\mathbf{\Sigma}_{\bar{D}} + \mathbf{\Sigma}_D)^{-1}\mu_{\bar{D}}\right\}^{1/2}\right].$$

Given values of $\mu_{\bar{D}}$, ρ_D, $(\sigma^2_{\bar{D}1}, \sigma^2_{\bar{D}2})$ and $\rho_{\bar{D}}$, we can explicitly calculate AUC_1 and AUC_M. Some values of AUC_1 and AUC_M are given in Table 1. What we find is that as the difference in correlations between biomarkers in the diseased and nondiseased populations increases, there are always gains in using a multivariate analysis relative to a univariate analysis. While AUC_M is always greater than AUC_1, based on the settings shown here, the mean has greater effect on the difference than the correlation. In limited settings not reported, we found the multivariate analysis to lead to even greater potential gains in classification accuracy when considering $p \geq 3$ biomarkers.

We next performed a simulation study in which a subset of the settings in Table 1 was considered. For simplification, $\sigma_{\bar{D}}$ was taken to be $(1, 1)$. The goal here was to assess the finite-sample properties of the proposed estimation procedures. We focused on estimation of AUC_M. We considered sample sizes $n = 100, 200$ and 500. For each setting, 500 simulation samples were generated. The proportion of diseased subjects was taken to be 50%. The estimation procedure in §3.1 was utilized. Variance estimation of AUC_M was done using

the bootstrap; 500 bootstrap simulations were performed within each step of the simulation. The simulation results are summarized in Table 2. We find that the procedure is practically unbiased for all sample sizes considered. While the standard error estimates based on the bootstrap tend to be slightly negatively biased for smaller sample sizes, this bias diminishes with larger sample sizes.

4.2. Cancer Biomarker Datasets

We consider two real-life applications of the proposed approach. The first is to data from Wieand et al. (1989). Two carbohydrate antigen proteins, CA125 and CA19-9, were measured on 90 subjects with pancreatic cancer and 51 controls free of disease. Thus, there are $p \equiv 2$ biomarkers. We expect both of these biomarkers to be higher in diseased individuals than in undiseased individuals. We fit the binormal model without covariates to these data. The results are summarized in Table 3. The estimated leave-one-out cross-validated curves for the univariate biomarkers, compared with that based on the estimated multivariate binormal model, are given in Figure 1. Based on the curve, we find that combining biomarkers leads to an increase in classification accuracy, although the gain is small relative to using the biomarker CA19-9 alone. For low values of the false positive rate (e.g. $0 - 0.2$), we find that the cross-validated ROC curve based on the multivariate binormal model yields better discrimination than either biomaker. The biomarker CA125 does not appear to discriminate between cases and controls well.

We now consider the data from a recent study reported by Etzioni et al. (1999). The data come from the Beta-Carotene and Retinol Efficacy Trial, a randomized trial that enrolled 12,025 men at elevated risk of lung cancer due to smoking or occupational exposure. While the primary disease outcome in the study was lung cancer, we focus on a subgroup in which prostate cancer was assessed. We consider data here on 71 subjects with prostate cancer and 68 control subjects. For these individuals, retrospective blood samples were available, so they were assayed for PSA (prostate-specific antigen). To simplify the discussion, we focus on the last sample before diagnosis for each subject so that each individual has only one measurement. In this situation, we focus on combining total PSA and the ratio of free to

total PSA, both of which have been suggested to have diagnostic utility in prostate cancer (Etzioni et al., 1999). It is assumed here that higher values of both correspond to greater risk of prostate cancer. Throughout this example, we have taken logarithmic transformations of both types of PSA measurements.

First, we compute the leave-one-out cross-validated ROC curves based on the unadjusted total PSA and PSA ratio; the plots are given in Figure 2. There, we find that while total PSA is a good discriminator of cases and controls, PSA ratio is not and that combining them leads to a deterioriation in classification accuracy. Next, we adjusted the two types of PSA measurements (corresponding to $\lambda = 0$ in (10) and (11) for age. Then the estimation procedure in §3.2 is applied to the residuals, and the multivariate binormal parameter estimates are computed. The numerical estimates of the multivariate binormal model parameters are given in Table 4; the cross-validated curves are given in Figure 3. Based on the plots, we find that there is not much discriminatory power using either total PSA or ratio of free to total PSA. Thus, the difference in classification accuracy in Figure 2 can be primarily attributed to the effect of age. Note that this analysis is much different than that done by Etzioni et al. (1999). There were interested in determining if longitudinal PSA profiles were capable of discriminating cases from controls. How to incorporate longitudinal data in the multivariate binormal proposed here remains a topic for future research.

5. Discussion

In this article, we have developed an approach to extending the binormal model for ROC curve estimation to accommodate multiple biomarkers. There is a lot of interest in new biomarkers found through gene expression and protein expression technologies (Sidransky, 2002); a pivotal issue then becomes how to combine information from multiple biomarkers. Our approach is complementary to previously described ones from the literature.

There are several potential extensions of the proposed methodology. In section 3.2.1, we treated λ, the transformation parameter, as a fixed constant. If we were to treat λ as unknown, then we could perform estimation in the model using a combination of the methods in Foster et al. (2001) and the Van der Warden correlation coefficient estimation procedure.

Our main motivation for treating λ as fixed is because such a model is easy to fit using standard statistical software packages, such as SAS, STATA or S-Plus.

While we have estimated correlations in the two-stage procedure in §3.2 using the Van der Warden procedure. We could formulate regression models for the correlations as well; this would correspond to a GEE2-type approach (Zhao and Prentice, 1990).

An advantage of the proposed approach is that it allows for incorporation of covariates in a very flexible manner. However, the effect of covariates on the biomarkers and their discriminatory power is multidimensional. An alternative approach would be to model covariate effects directly on the ROC curve (Pepe, 2000). One could then envision marginal models for ROC curves for multiple biomarkers. This is an area that needs further study.

Acknowledgments

The author thanks Tim Johnson and Zheng Yuan for useful discussions.

Appendix

Justification of two-stage estimation procedures

To make results concrete, we consider the case where $p = 2$ and consider only the diseased population. Suppose we consider the following joint model:

$$h_1(Y) = \mathbf{X}^T\beta_{D1} + \epsilon_1$$
$$h_2(Y) = \mathbf{X}^T\beta_{D2} + \epsilon_2$$

where h_j is a monotonic function ($j = 1, 2$) and (ϵ_1, ϵ_2) is distributed with marginal distribution functions F_1 and F_2 and copula function (7). Define $\Psi_j(y, z) = h_j(y) - z^T\beta_{Dj}$, $j = 1, 2$. At the true h_j and β_{Dj} ($j = 1, 2$), $\Psi_j(Y, \mathbf{X})$ does not depend on the distribution of (ϵ_1, ϵ_2). This fact was also utilized by Fine and Jiang (2000) in a survival analysis setting. For the model (10), this means that $(Y_j^\lambda - 1)/\lambda - \mathbf{X}^T\beta_{Dj}$ ($j = 1, 2$) does not depend on the joint distribution of (ϵ_1, ϵ_2) and similarly for (11).

References

Baker, S. (2000). Identifying combinations of cancer markers for further study as triggers of early intervention. *Biometrics* **56**, 1082 – 1087.

Clayton, D. G. (1978). A model for association in bivariate life tables and its application in epidemiological studies of familial tendency in chronic disease incidence. *Biometrika* **65**, 141 – 151.

Dale, J. R. (1986). Global cross-ratio models for bivariate, discrete, ordered responses. *Biometrics* **42**, 909–917.

Etzioni, R., Pepe, M., Longton, G., Hu, C., and Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. *Medical Decision Making* **19**, 242 – 251.

Fine, J. P. and Jiang, H. (2000). On association in a copula with time transformations. *Biometrika* **87**, 559 – 571.

Foster, A., Tian, L. and Wei, L. J. (2001). Estimation for the Box-Cox transformation model with assuming parametric error distribution. *J. Am. Statist. Assoc.* **96**, 1097 – 1101.

Genest, C., and Rivest, L.-P. (1993). Statistical inference procedures for bivariate Archimedean copulas. *J. Am. Statist. Assoc.* **88**, 1034 – 1043.

Hanley, J. A. (1988). The robustness of the 'binormal' assumptions used in fitting ROC curves. *Medical Decision Making* **8**, 197 – 203.

Heagerty, P. J., and Zeger, S. L. (1996). Marginal regression models for clustered ordinal measurements. *J. Am. Statist. Assoc.*, **91**, 1024–1036.

Hsu, L. and Prentice, R. L. (1996). On assessing the strength of dependency between failure time variates. *Biometrika* **83**, 491 – 506.

Kawinski, E., Levine, E. and Chadha, K. (2002). Thiophilic interaction chromatography facilitates detection of various molecular complexes of prostate-specific antigen in fluids. *Prostate*, **50**, 145 – 153.

Joe, H. (1997). *Multivariate Models and Dependence Concepts.* London: Chapman and Hall.

Klaassen, C. A. J., and Wellner, J. A. (1997). Efficient estimation in the bivariate normal copula model: normal margins are least-favorable. *Bernoulli*, 55 – 77.

Lin, Y. and Jeon, Y. (2003). Discriminant analysis through a semiparametric model. *Biometrika* **90**, 379 – 392.

McIntosh, M. W. and Pepe, M. S. (2002). Combining several screening tests: optimality of the risk score. *Biometrics* **58**, 657 – 664.

Molenberghs, G., and Lesaffre, E. (1994). Marginal modelling of correlated ordinal data using a multivariate Plackett distribution. *J. Am. Statist. Assoc.* **89**, 633-644.

Nelsen, R. (1999). *An Introduction to Copulas.* New York: Springer.

Oakes, D. (1986). Semiparametric inference in a model for association in bivariate survival data. *Biometrika* **73** 353 – 361.

Oakes, D. (1989). Bivariate survival models induced by frailties. *J. Am. Statist. Assoc.* **84**, 487 – 493.

Pepe, M. S. (2000). An interpretation for the ROC curve and inference using GLM procedures. *Biometrics* **84**, 352 – 359.

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

Pepe, M. S. and Thompson, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123 – 140.

Sidransky, D. (2002). Emerging molecular markers of cancer. *Nature Reviews Cancer* **2**, 210 – 219.

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. (1986). Indices of discrimination or diagnostic accuracy: their ROCs and implied models. *Psychological Bulletin* **99**, 100 – 117.

Van der Vaart, A. (2000). *Asymptotic Statistics*. Cambridge: Cambridge University Press.

Weinstein, M. C., Fineberg, H. V., Elstein, A. S., Frazier, N. S., Neuhauser, D., Neutra, R. R., and McNeil, B. J. (1980). *Clinical Decision Analysis*. Philadelphia: W. B. Saunders.

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 – 92.

Zhao LP, Prentice R. Correlated binary regression using a quadratic exponential model. Biometrika l990; 77: 642-48.

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

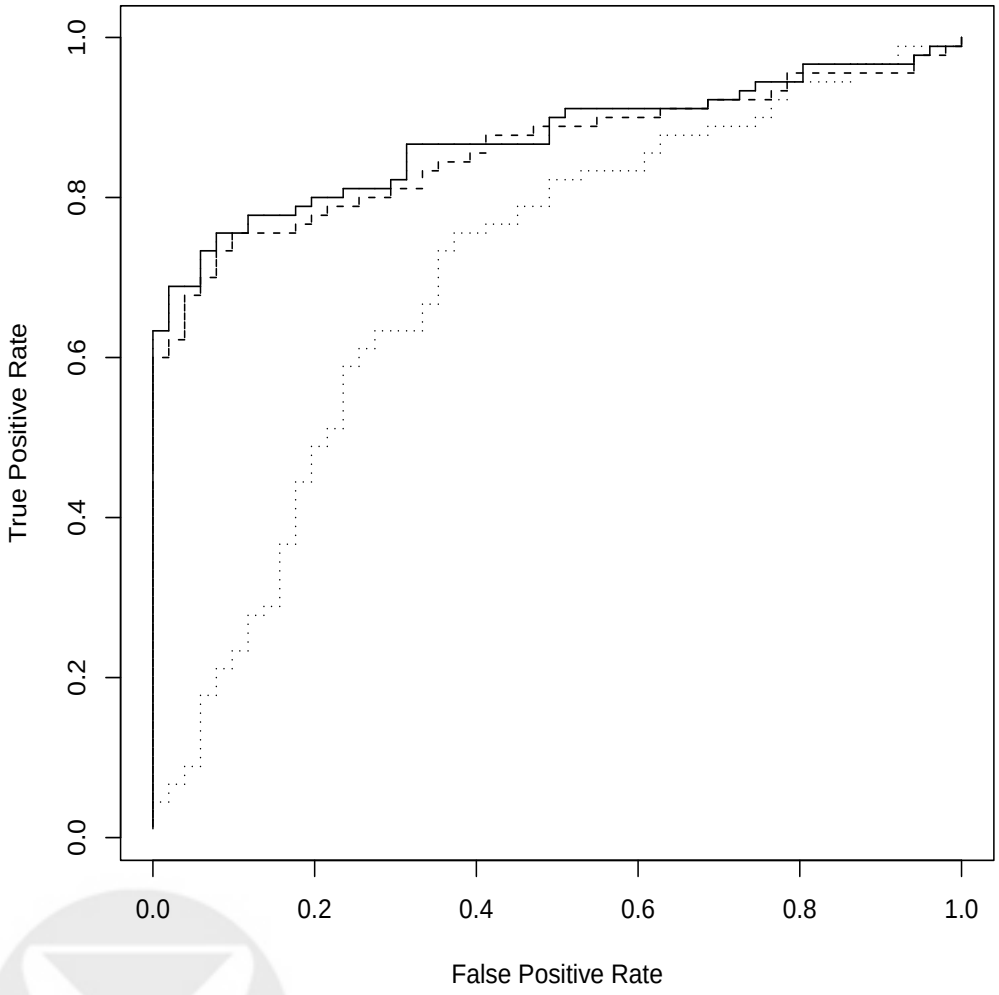

Figure 1: Comparison of cross-validated ROC curves for combined and univariate biomarkers (CA125 and CA19-9) for pancreatic cancer data. The solid line represents the ROC from the multivariate binormal model estimation procedure in §??. The dashed line is the ROC curve for CA19-9, while the dotted line is for CA125.

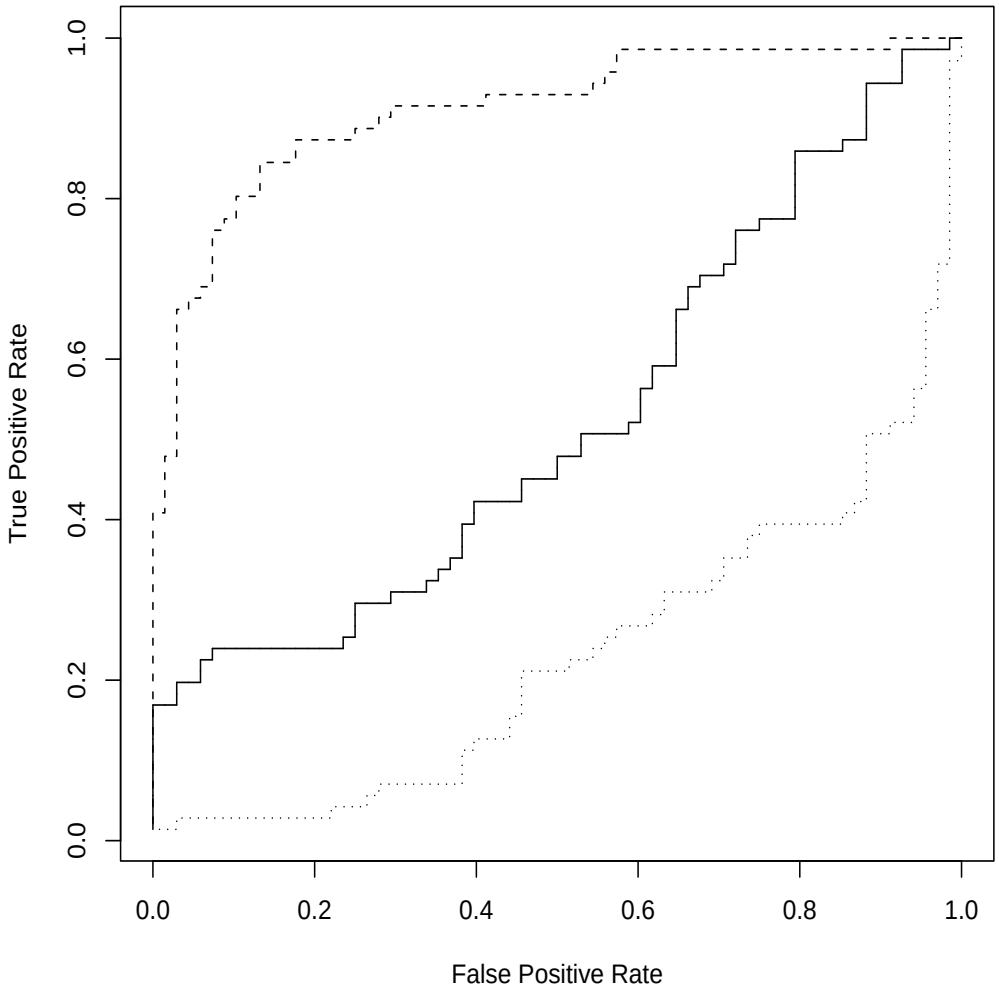

Figure 2: Cross-validated ROC curves, using multivariate binormal model (solid line) and univariate binormal model for log(free PSA) (dashed line) and log(free PSA/total PSA) (dashed line).

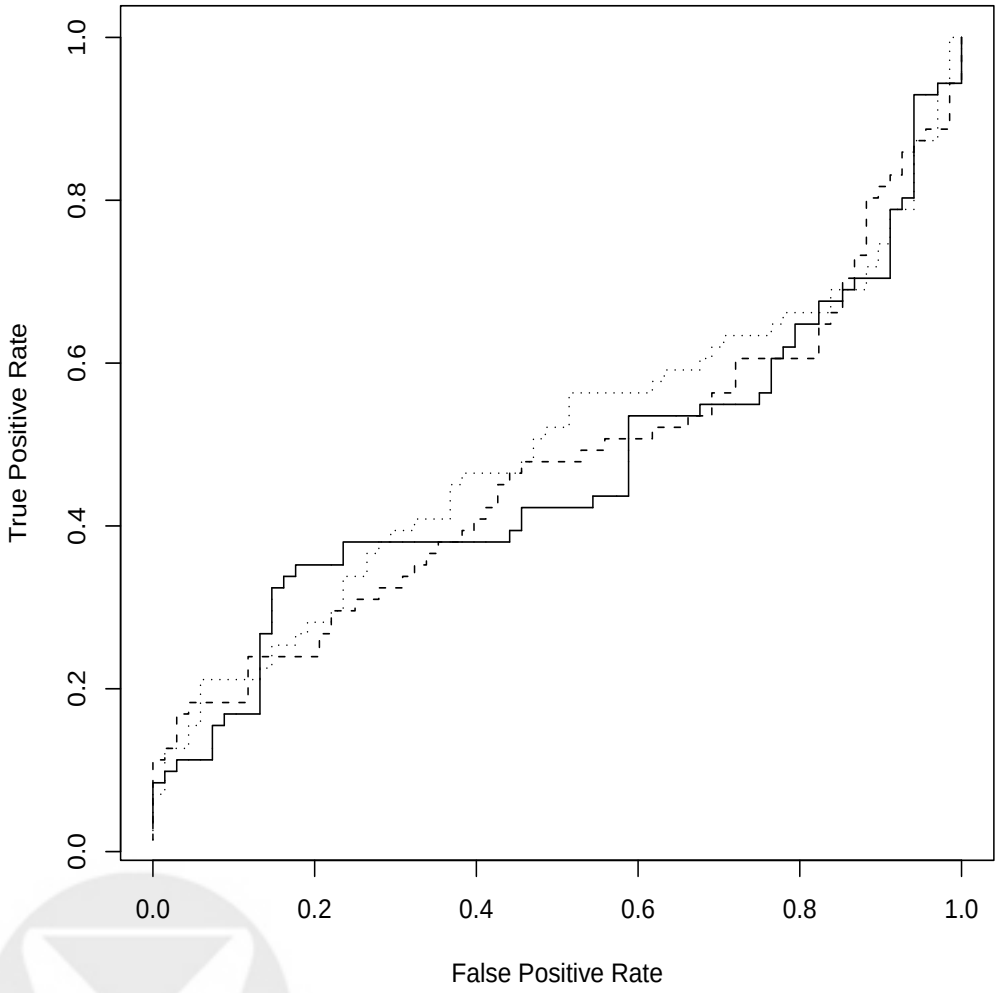

Figure 3: Cross-validated ROC curves, adjusting for age, using multivariate binormal model (solid line) and univariate binormal model for log(free PSA) (dashed line) and log(free PSA/total PSA) (dashed line).

Table 1. *Analytical comparisons of AUC_1 and AUC_M*

$\mu_{\bar{D}}$	$\sigma_{\bar{D}}$	$\rho_{\bar{D}}$	ρ_D	AUC_1	AUC_M
$(-0.5, -0.5)$	$(1, 1)$	0.2	0.2	0.638	0.676
$(-0.5, -0.5)$	$(1, 1)$	0.3	0.2	0.638	0.673
$(-0.5, -0.5)$	$(1, 1)$	0.2	0.8	0.638	0.658
$(-1.5, -0.5)$	$(1, 1)$	0.2	0.2	0.856	0.858
$(-1.5, -0.5)$	$(1, 1)$	0.3	0.2	0.856	0.856
$(-1.5, -0.5)$	$(1, 1)$	0.2	0.8	0.856	0.859
$(-0.5, -1.5)$	$(1, 1)$	0.2	0.2	0.638	0.857
$(-0.5, -1.5)$	$(1, 1)$	0.3	0.2	0.638	0.856
$(-0.5, -1.5)$	$(1, 1)$	0.2	0.8	0.638	0.860
$(-0.5, -0.5)$	$(1, 0.5)$	0.2	0.2	0.638	0.700
$(-0.5, -0.5)$	$(1, 0.5)$	0.3	0.2	0.638	0.698
$(-0.5, -0.5)$	$(1, 0.5)$	0.2	0.8	0.638	0.678
$(-1.5, -0.5)$	$(1, 0.5)$	0.2	0.2	0.856	0.862
$(-1.5, -0.5)$	$(1, 0.5)$	0.3	0.2	0.856	0.861
$(-1.5, -0.5)$	$(1, 0.5)$	0.2	0.8	0.856	0.859
$(-0.5, -1.5)$	$(1, 0.5)$	0.2	0.2	0.638	0.911
$(-0.5, -1.5)$	$(1, 0.5)$	0.3	0.2	0.638	0.910
$(-0.5, -1.5)$	$(1, 0.5)$	0.2	0.8	0.638	0.924

Table 2. *Summary of simulation results for AUC_M*

n	$\mu_{\bar{D}}$	$\rho_{\bar{D}}$	ρ_D	bias($\widehat{AUC}_M$)	SE($\widehat{AUC}_M$)	SEE($\widehat{AUC}_M$)
100	$(-0.5, -0.5)$	0.2	0.2	0.01	0.220	0.201
	$(-1.5, -0.5)$	0.3	0.2	0.00	0.207	0.189
	$(-0.5, -1.5)$	0.2	0.8	-0.01	0.211	0.193
200	$(-0.5, -0.5)$	0.2	0.2	0.01	0.155	0.148
	$(-1.5, -0.5)$	0.3	0.2	0.00	0.147	0.141
	$(-0.5, -1.5)$	0.2	0.8	0.00	0.159	0.149
500	$(-0.5, -0.5)$	0.2	0.2	-0.01	0.098	0.096
	$(-1.5, -0.5)$	0.3	0.2	-0.01	0.093	0.091
	$(-0.5, -1.5)$	0.2	0.8	-0.01	0.094	0.092

Note: SE($\widehat{AUC}_M$) is the empirical standard error of $\widehat{AUC}_M$, while SEE($\widehat{AUC}_M$) is the estimated standard error based on bootstrap distribution of $\widehat{AUC}_M$.

Table 3. *Multivariate binormal parameter estimates for Wieand et al. (1989) pancreatic cancer data*

Biomarker	μ_U	σ_U	ρ_U	ρ_D
CA19-9	-1.199	0.437	-0.121	0.134
CA125	-0.737	0.992		

Note: ρ_U denotes correlation between CA19-9 and CA-125 in undiseased population; ρ_D denotes correlation between CA19-9 and CA-125 in diseased population.

Table 4. *Multivariate binormal parameter estimates for Etzioni et al. (1999) prostate cancer data*

Biomarker	μ_U	σ_U	ρ_U	ρ_D
total PSA	0.076	0.612	-0.430	-0.505
PSA ratio	0.010	0.594		

Note: Both markers have been adjusted for age. PSA ratio is ratio of free to total PSA; ρ_U denotes correlation between total PSA and PSA ratio in undiseased population; ρ_D denotes correlation between total PSA and PSA ratio in diseased population.