Journal of Data Science **6**(2008), 1-13

Estimating Optimum Linear Combination of Multiple Correlated Diagnostic Tests at a Fixed Specificity with Receiver Operating Characteristic Curves

Feng Gao[1], Chengjie Xiong[1], Yan Yan[1], Kai Yu[2] and Zhengjun Zhang[3]
[1] *Washington University in St. Louis,* [2] *National Cancer Institute and*
[3] *University of Wisconsin at Madison*

Abstract: Receiver operating characteristic (ROC) methodology is widely used to evaluate diagnostic tests. It is not uncommon in medical practice that multiple diagnostic tests are applied to the same study sample. A variety of methods have been proposed to combine such potentially correlated tests to increase the diagnostic accuracy. Usually the optimum combination is searched based on the area under a ROC curve (AUC), an overall summary statistics that measures the distance between the distributions of diseased and non-diseased populations. For many clinical practitioners, however, a more relevant question of interest may be ”what the sensitivity would be for a given specificity (say, 90%) or what the specificity would be for a given sensitivity?”. Generally there is no unique linear combination superior to all others over the entire range of specificities or sensitivities. Under the framework of a ROC curve, in this paper we presented a method to estimate an optimum linear combination maximizing sensitivity at a fixed specificity while assuming a multivariate normal distribution in diagnostic tests. The method was applied to a real-world study where the accuracy of two biomarkers was evaluated in the diagnosis of pancreatic cancer. The performance of the method was also evaluated by simulation studies.

Key words: Diagnostic accuracy, optimal linear combination test, receiver operating characteristic (ROC) curve, sensitivity, specificity.

1. Introduction

In recent years many efforts have been made on studying biomarkers that could provide accurate and non-invasive ways of disease diagnosis or prognosis. Many of these biomarkers are measured in a continuous scale and receiver operating characteristic (ROC) curve is widely used for evaluating the accuracy of such a continuous diagnostic test (Hanley and McNeil 1984; Hanley 1989; Begg 1991). Suppose that, based on some gold standard independent of the diagnostic tests to be evaluated, subjects belong to 1 of 2 basic conditions – diseased (D^+)

and non- diseased (D^-). The ROC curve evaluates the ability of the diagnostic test to discriminate the two conditions. By plotting the true positive rates (sensitivity) versus the false positive rates (1-specificity) across all possible thresholds, ROC curve reflects the relative trade-off between true and false positive rates. The area under the ROC curve (AUC) measures the distance between the distributions of diseased and non-diseased populations and is frequently used as a global measure for the accuracy of the diagnostic test (Swets and Pickett 1982; DeLong, Vernon and Bollinger 1985; Ma and Hall 1993). If a test could perfectly discriminate, then there exists a cut-point above which all member of one group (diseased or non-diseased) will fall and below which all members of the alternative group will fall. The ROC curve would then pass through the point (0,1) on the grid $[0, 1] \times [0, 1]$, with an AUC of one. The closer the AUC comes to this ideal, the more discriminating ability the test has. Zhou et al (2002) and Pepe (2003) provide excellent reviews of the existing methods on the analysis of ROC curves.

In clinical studies, it is not uncommon that multiple diagnostic tests are applied to the same sample. In such a case, the diagnostic tests are more likely to be correlated. A variety of methods have been proposed to evaluate and compare the performance of such correlated diagnostic tests. Greenhouse and Mantel (1950), and Linnet (1987) compared two sensitivities at a single fixed specificity. McClish (1987) proposed a way to assess the relative diagnostic accuracy of independent ROC curves using the difference of areas under curves (AUC). Metz *et al.* (1984) generalized the statistical comparison of the binormal ROC model (i.e., assuming the data in both diseased and non-diseased groups follow normal distributions) to bivariate case for comparing the difference on AUC between correlated ROC curves. DeLong *et al.* (1988) and Venkatraman and Begg (1996) developed non-parametric methods to compare the areas under two ROC curves. Wieand and colleagues (1989) proposed a more general family of non-parametric statistics to compare the weighted average of sensitivities. Their method can be used to compare the diagnostic tests either over a restricted range of specificity or under an entire ROC curve. However, since different markers are usually representative to different aspects of diseases, it is desirable to combine the correlated tests to increase the diagnostic accuracy. Assuming that the biomarkers of interest have a multivariate normal distribution in each of the diseased and non-diseased populations, Su and Liu (1993) worked under a linear discriminant analysis framework to separate the two conditions. They showed that the linear combination derived from discriminant function maximizes the area under the ROC curve. Based on the same binormal assumptions, recently Xiong and colleagues (2004) proposed an approach to construct the optimum linear combination over all possible linear combinations under a ROC analysis framework. Based on the eigenvalue of the optimum linear combination of the diagnostic tests, they presented closed forms for the estimation of maximum AUC and its variance. Both of the above methods developed the optimum linear combinations based on the area under a ROC curve, an overall summary statistics that measures the distance between the distributions of diseased and non-diseased populations. In clinical applications, however, a marker's usefulness is generally determined by its specific settings. For example, a test with 20% false positive rate (80% specificity) may be acceptable for cancer prognosis, but usually will be too high for cancer screening. Therefore, a more frequently question raised by a clinician could be "How much sensitivity (specificity) can be achieved at a given specificity (sensitivity)?".

This article addresses the problem of combining multiple correlated diagnostic tests under a similar framework as Xiong *et al.* (2004). Instead of searching optimal linear combination that maximizes AUC, our method is seeking an optimum linear combination in discriminating between the diseased population and the healthy population at a single fixed specificity. More specifically, we consider all possible linear combinations of multiple diagnostic tests and numerically search for the best set of coefficients (weights) that maximizes the sensitivity at a given specificity of interest. The standard deviation and 95% confidence interval of the estimate are constructed taking a parametric bootstrap approach (Gentle 2002). The method is exemplified with a study on the diagnosis of pancreatic cancer where two serum markers are measured at 90 patients with pancreatic cancer and 51 patients with pancreatitis. The performance of the method is also evaluated by simulation studies.

2. Method

We assume that a total of r tests are used for each subject in both the diseased population and the healthy population. Without loss of generality, we assume that higher values of each test are associated with the positive results. Let D$^+$ and D$^-$ denote the diseased (i.e., the positive condition) group and the non-diseased (i.e., the negative condition) group respectively. Let $\mathbf{X} = (X^1, X^2, \ldots, X^r)^t$ (t stands for the transpose) be the values of the r test results for a subject in group D$^+$, and $\mathbf{Y} = (Y^1, Y^2, \ldots, Y^r)^t$ be the values of the r test results for a subject in group D$^=$. We assume that $(X^1, X^2, \ldots, X^r)^t$ follows a multivariate normal distribution $MVN_r(\boldsymbol{\mu}^+, \Sigma^+)$ with mean vector $\boldsymbol{\mu}^+ = (\mu_1^+, \ldots, \mu_r^+)^t$ and covariance matrix $\Sigma^+ = (\sigma_{ij}^+)_{1\leq I,j\leq r}$ and that $(Y^1, Y^2, \ldots, Y^r)^t$ follows another multivariate normal distribution $MVN_r(\boldsymbol{\mu}^-, \Sigma^-)$ with mean vector $\boldsymbol{\mu}^- = (\mu_1^-, \mu_2^-, \ldots, \mu_r^-)^t$ and covariance matrix $\Sigma^- = (\sigma_{ij}^-)_{1\leq I,j\leq r}$. As mentioned earlier, we assume that $\boldsymbol{\mu}^+ > \boldsymbol{\mu}^-$ in each test. We also assume that Σ^+ and Σ^- are positive definite. Considering the scenario of a single test (i.e., i-th test), let (μ_i^+, σ_{ii}^+) and (μ_i^-, σ_{ii}^-)

denote means and variances in the diseased and non-diseased groups respectively. For a given specificity Q, the cut-off value C_γ in the non-diseased group can be determined as $C_\gamma = \mu_i^- + \sqrt{\sigma_{ii}^-}\Phi^{-1}(Q)$. After applying C_γ to the diseased group, the corresponding sensitivity will be

$$1 - \Phi\left(\frac{C_\gamma - \mu_i^+}{\sqrt{\sigma_{ii}^+}}\right) = \Phi\left(\frac{C_{\mu_i}^+ - C_\gamma}{\sqrt{\sigma_{ii}^+}}\right) = \Phi\left(\frac{\mu_i^+ - \mu_i^- - \sqrt{\sigma_{ii}^-}\Phi^{-1}(Q)}{\sqrt{\sigma_{ii}^+}}\right).$$

Therefore, the binormal ROC model can be written as

$$f(Q) = \Phi[a + b\Phi^{-1}(1-Q)], \quad \text{with } a = \frac{\mu_i^+ - \mu_i^-}{\sqrt{\sigma_{ii}^+}} \text{ and } b = \frac{\sigma_{ii}^-}{\sigma_{ii}^+}, \tag{2.1}$$

where the double subscripted σ represents variances, $\Phi(\cdot)$ is the cumulative distribution of a standard normal distribution, $\Phi^{-1}(\cdot)$ is its inverse function, and Q is a given specificity. Note that in our notation σ represents variances rather than standard deviations. The above model plays a central role in ROC analysis similar to the role of normal distribution in classical statistical modeling, and it has been shown that this model provides a good approximation to a wide range of ROC curves encountered in practice (Pepe 2003; Hanley 1996). In the presence of multiple (correlated) tests, we seek a linear combination of r diagnostic tests such that the sensitivity is maximized over all possible linear combinations when the specificity is fixed at Q (preferably $0.5 < Q < 1$). Let $\mathbf{w} = (w_1, w_2, \ldots, w_r)^t$ be a set of weights (coefficients), $S = \mathbf{w}^t X$ and $T = \mathbf{w}^t \mathbf{Y}$ be the scores of linear combinations of the r diagnostic tests at the diseased and health populations respectively. The corresponding ROC associated with S and T is given by,

$$g(Q) = \Phi[c + d\Phi^{-1}(1-Q)], \tag{2.2}$$

where

$$\begin{aligned} c &= \frac{\mathbf{w}^t - \boldsymbol{\mu}^+ - \mathbf{w}\boldsymbol{\mu}^-}{\mathbf{w}^t\Sigma^+\mathbf{w}} \\ d &= \sqrt{\frac{\mathbf{w}^t\Sigma^-\mathbf{w}}{\mathbf{w}^t\Sigma^+\mathbf{w}}}. \end{aligned}$$

Since the cumulative distribution of the standard normal distribution Φ is strictly monotonic, the maximization of $g(Q)$ given Q over the choice of $\mathbf{w}$ is equivalent to the maximization of,

$$C(\mathbf{w}) = c + d\Phi^{-1}(1-Q) = \frac{\mathbf{w}^t\boldsymbol{\mu}^+ - \mathbf{w}^t\boldsymbol{\mu}^- + \sqrt{\mathbf{w}^t\Sigma^-\mathbf{w}\Phi^{-1}(1-Q)}}{\sqrt{\mathbf{w}^t\Sigma^-\mathbf{w}}},$$

where $\mathbf{w} = (w_1, w_2, \ldots, w_r)^t$ is obtained numerically with the constraint $\sum_{i=1}^r w_i^2$ $= 1$.

Since the distribution of the maximal sensitivity, $g(Q)$, is analytically intractable, the standard deviation and confidence interval of the estimated sensitivity will be constructed taking a parametric bootstrap approach. Specifically, means and covariance matrices will be estimated from the study sample, and 1000 samples (each with the same number of observations as in the original data) will be generated from multivariate normal distributions based on these estimates. For each of the generated samples, a maximal sensitivity, $g(Q)^*$, will be estimated based on (2.2). Then, the standard deviation and 95% confidence interval of $g(Q)$ can be obtained based on the estimated distribution of $g(Q)^*$ (Gentle 2002).

3. Application: Biomarkers for the Diagnosis of Pancreatic Cancer

For illustration, we apply the proceeding method to a real-world data on the diagnosis of pancreatic cancer with two tumor markers (CA19-9 and CA125). CA19-9 is a carbohydrate antigen that tends to be elevated especially in subjects with carcinomas of the gastrointestinal tract while CA125 is a cancer antigen that is associated with a variety of malignancies including breast, cervix, pancreas, and lung, etc. A study conducted at Mayo Clinic considered 90 “cases” of patients with pancreatic cancer (D^+) and 51 ”controls” of patients with pancreatitis (D^-). Serum CA19-9 and CA125 were measured on each of these patients and both of the markers were measured in continuous scales. The data was first presented by Wieand *et al.* (1989) to compare the relative accuracy of the two biomarkers for the diagnosis of pancreatic cancer. Zhou and others (2002) used the data to illustrate the maximum likelihood method and more recently Cai and Moskowitz1 (2004) exemplified the data with two semi-parametric approaches for fitting ROC models. The objective of our current analysis is to derive an optimum linear combination of the two markers that maximizes the sensitivity over all possible linear combinations at a fixed specificity (90%, say).

Let $\mathbf{X} = (X^1, X^2)^t$ be the values of CA19-9 and CA125 for pancreatic cancer patients and $\mathbf{Y} = (Y^1, Y^2)^t$ be the marker values for pancreatitis patients respectively. The original distributions of CA19-9 and CA125 are found to be badly skewed to the right because some of the marker values tend to be extremely large, and thus a logarithm transformation was performed for both markers to improve the normality. Based on the behavior of the majority data (Figure 1), we can assume that $\log(\mathbf{X})$ and $\log(\mathbf{Y})$ have a bivariate normal distribution,

$$\log(\mathbf{X}) \sim MVN(\boldsymbol{\mu}^+, \Sigma^+) \quad \text{and} \quad \log(\mathbf{Y}) \sim MVN(\boldsymbol{\mu}^-, \Sigma^-).$$

Then the maximum likelihood estimates of the parameters in the cancer group can be obtained as

$$\hat{\boldsymbol{\mu}}^{+} = (5.42, 3.26)^{t} \quad \text{and} \quad \hat{\Sigma}^{+} = \begin{pmatrix} 5.483 & 0.328 \\ 0.328 & 0.977 \end{pmatrix}$$

and the corresponding estimates in the pancreatitis group are

$$\hat{\boldsymbol{\mu}}^{-} = (2.47, 2.67)^{6} \quad \text{and} \quad \hat{\Sigma}^{-} = \begin{pmatrix} 0.748 & -0.095 \\ -0.095 & 0.612 \end{pmatrix}.$$

Without considering the possible correlations between the two markers, approximately 78% sensitivity can be achieved for CA19-9 alone while the maximum sensitivity for CA125 alone is 34% at a given 90% specificity. After applying our proposed method, the optimum weights are searched numerically as 0.89 for CA19-9 and 0.455 for CA125. For a fixed 90% specificity, the resulting linear combination, $0.89 \times \log(CA19-9) + 0.455 \times \log(CA125)$, will achieve an approximately 80% (SD=4.0%) sensitivity , with a 95% confidence interval of $[72.0\%, 87.4\%]$. The dotted line in Figure 2 corresponds to the 90% specificity of the resultant optimum linear combination which is the best one over all possible linear combinations, $w_1 \times \log(CA19-9) + w_2 \times \log(CA125)$, such that $w_1^2 + w_2^2 = 1$.

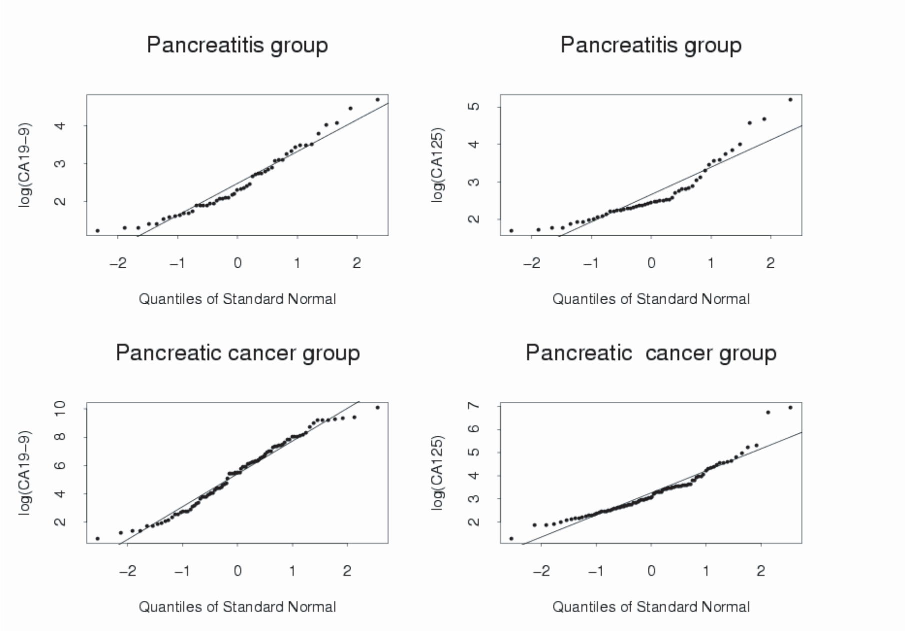

Figure 1: The Q-Q plots for $\log(CA19-9)$ and $\log(CA125)$ in 51 pancreatitis patients (D^{-}) and 90 pancreatic cancer patients (D^{+}).

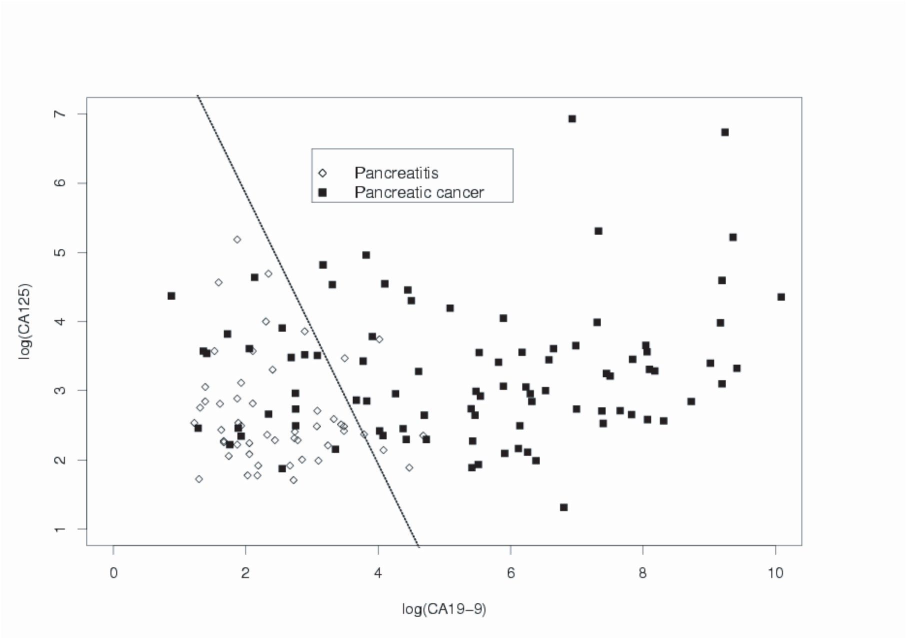

Figure 2: The scatter plot for $\log(CA19-9)$ versus $\log(CA125)$in 51 pancreatitis patients (D^-) and 90 pancreatic cancer patients (D^+), where dotted line corresponds the optimum linear combination that maximizes sensitivity at a fixed 90% specificity.

4. Simulation Studies

Simulation studies are designed to evaluate the performance of proposed method in the presence of correlated multiple diagnostic tests. In practice, the true mean and the true covariance matrix of a vector of multivariate diagnostic tests are rarely known, and the best linear combination has to be derived based on the estimated means and covariance matrices. Therefore, it is important to assess how the sample size and inter-marker correlation affect the performance of the estimated optimum combination. The simulation assumes 3 correlated diagnostic tests. These diagnostic tests in the diseased (D^+) group are assumed to have a 3-dimensional normal distribution $MVN(\boldsymbol{\mu}^+, \Sigma^+)$ of

$$\boldsymbol{\mu}^+ = (2.5, 4.5, 6.0)^t \quad \text{and} \quad \Sigma^+ = \begin{pmatrix} \sigma_1^+ & \rho_1\sqrt{\sigma_1^+\sigma_2^+} & \rho_1\sqrt{\sigma_1^+\sigma_3^+} \\ \rho_1\sqrt{\sigma_1^+\sigma_2^+} & \sigma_2^+ & \rho_1\sqrt{\sigma_2^+\sigma_3^+} \\ \rho_1\sqrt{\sigma_1^+\sigma_3^+} & \rho_1\sqrt{\sigma_2^+\sigma_3^+} & \sigma_3^+ \end{pmatrix}$$

with the vector of variance to be $(\sigma_1^+, \sigma_2^+, \sigma_3^+) = (3, 2, 1)$. The tests in the healthy (D^-) group are also assumed a 3-dimensional normal distribution $MVN(\boldsymbol{\mu}^-, \Sigma^-)$

of

$$\boldsymbol{\mu}^- = (2.0, 3.0, 4.0)^t \quad \text{and} \quad \Sigma^- = \begin{pmatrix} \sigma_1^- & \rho_1\sqrt{\sigma_1^-\sigma_2^-} & \rho_1\sqrt{\sigma_1^-\sigma_3^-} \\ \rho_1\sqrt{\sigma_1^-\sigma_2^-} & \sigma_2^- & \rho_1\sqrt{\sigma_2^-\sigma_3^-} \\ \rho_1\sqrt{\sigma_1^-\sigma_3^-} & \rho_1\sqrt{\sigma_2^-\sigma_3^-} & \sigma_3^- \end{pmatrix},$$

with the variance vector of $(\sigma_1^-, \sigma_2^-, \sigma_3^-) = (6, 2, 4)$.

For simplicity, we consider a common correlation parameter ($\rho = \rho^+ = \rho^-$ in our simulation and let ρ take 3 values ($\rho = 0.2, 0.5, 0.8$). We also assume that diseased and healthy groups have an equal sample size, and in the simulations 4 sample sizes ($N =$ 25, 50, 100 and 200) are considered for each group. For each selected sample size, 1000 random samples are generated from $MVN_3(\boldsymbol{\mu}^+, \Sigma^+)$ and $MVN_3(\boldsymbol{\mu}^-, \Sigma^-)$ at a given ρ respectively. In this study the simulation was implemented by the statistical package S-Plus (version 6.2). The random samples were generated from the function RMVNORM (the random generation function for the multivariate normal distribution) while the optimum weights (coefficients) $\mathbf{w} = (w_1, w_2, w_3)^t$ for linear combinations were searched numerically by the function NLMINB (the function for nonlinear minimizations subject to box constraint). To satisfy the constraint of $\sum_{i=1}^3 w_i^2 = 1$, the actual minimization was performed on the unconstrained parameters $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_r)^t$ such that $w_i = \gamma_i / \sqrt{\sum \gamma_i^2}$. In the simulations, we evaluated the performance of the method given $Q = 80\%$ and 90%, two specificities that are usually of most interest to clinicians.

By assuming that all the mean vectors and variance-covariance matrices in both diseased and healthy populations are known, Table 1 shows the optimum weights and the expected maximum sensitivities at different combinations of ρ and Q. These optimum weights will produce the best linear combination that gives the maximum sensitivity over all possible linear combinations of the 3 diagnostic tests. The results show that the optimum weights (and thus the maximum sensitivity) are a function of the inter-marker correlation. When there exists a weak correlation among these 3 biomarkers ($\rho = 0.2$) at a given $Q = 90\%$, for example, the optimum weights $\hat{\mathbf{w}} = (-0.079, 0.820, 0.567)^t$ will give the best linear combination as $S = \sum_{i=1}^3 \hat{w}_i X_i$ in the diseased sample and $T = \sum_{i=1}^3 \hat{w}_i Y_i$ in the healthy sample. Then a maximum 52% sensitivity can be achieved based on the scores of S and T. In contrast, a different set of optimum weights $\hat{\mathbf{w}} = (-0.589, 0.521, 0.617)$ will be obtained in the presence of a strong correlation ($\rho = 0.8$) among the 3 tests, and the resultant combined test will allow us a maximum 72% sensitivity at the fixed $Q = 90\%$. The optimum weights in Table 1 are consistent in signs with the expected weights $\hat{\mathbf{w}} = (-0.3845, 0.6767, 0.1692)$ by Xiong *et al.* (2004) who took a similar parameter setups as ours but searched for the optimum combination maximizing area under ROC curve (AUC). Our simulations show that optimum weights are also a function of the specificity (Q). In the presence of a weak inter-marker correlation ($\rho = 0.2$), for example, the optimum coefficients for $Q = 80\%$ are $\hat{\mathbf{w}} = (-0.080, 0.626, 0.776)$ while the coefficients are $\hat{\mathbf{w}} = (-0.079, 0.820, 0.567)$ for $Q = 90\%$, and the corresponding linear combinations will result in 75% and 52% maximum sensitivities respectively. Our finding is consistent to the work by Anderson and Bahadur (1962) that generally there is no unique linear combination superior to all others over the entire range of specificities (sensitivities).

Table 1: The optimum weights and the expected maximum sensitivity at a fixed specificity (Q) when the means and variance-covariance matrices of the diagnostics tests are known, where ρ represents the inter-test correlation.

Q	ρ	Optimum weights			Expected Maximum Sensitivity
		w_1	w_2	w_3	
80%	0.2	-0.080	0.626	0.776	0.751
	0.5	-0.354	0.532	0.770	0.752
	0.8	-0.551	0.308	0.776	0.902
90%	0.2	-0.079	0.820	0.567	0.524
	0.5	-0.342	0.782	0.521	0.529
	0.8	-0.589	0.521	0.617	0.721

In real-world applications, the true mean and the true covariance matrix of a vector of multivariate diagnostic tests are rarely known, and the best linear combination of the diagnostic tests has to be derived based on the estimated means and covariance matrices. Table 2 presents the averages of estimated maximum sensitivity and its standard deviation based on 1000 random samples. The results show that the estimated maximum sensitivity becomes closer to the expected ones as the sample size increases and an accurate estimate can be achieved even in a relatively small sample size. The last column in Table 2 shows the empirical coverage probabilities of 95% confidence interval (CI). We see that, though the empirical coverage probabilities tend to be lower than the nominal 95% coverage probability when sample sizes are relatively small, the estimated confidence intervals perform very well for moderate to large sample sizes.

Table 2: The averages of the estimated maximum sensitivity, the average of the estimated standard deviation (SD), and the empirical coverage of 95% confidence intervals based on 1000 random samples, where ρ represents inter-test correlation and N is the sample size in each group.

$Q = 80\%$				$Q = 90\%$			
ρ	N	$\hat{g}(Q)\pm$ SD	CI	ρ	N	$\hat{g}(Q)\pm$ SD	CI
0.2	25	0.777±0.114	0.928	0.2	25	0.581±0.141	0.921
	50	0.765±0.086	0.932		50	0.555±0.106	0.933
	100	0.762±0.062	0.942		100	0.547±0.077	0.943
	200	0.755±0.046	0.942		200	0.535±0.055	0.947
0.5	25	0.776±0.116	0.923	0.5	25	0.582±0.141	0.922
	50	0.767±0.088	0.935		50	0.551±0.105	0.923
	100	0.764±0.064	0.948		100	0.540±0.077	0.950
	200	0.755±0.047	0.951		200	0.531±0.055	0.956
0.8	25	0.904±0.081	0.931	0.8	25	0.752±0.127	0.916
	50	0.905±0.057	0.935		50	0.736±0.098	0.934
	100	0.906±0.041	0.939		100	0.732±0.072	0.946
	200	0.903±0.030	0.952		200	0.725±0.052	0.949

5. Discussion

In this paper, we proposed an approach to estimate the maximum sensitivity at a fixed specificity in the presence of multiple correlated diagnostic tests. Although we focused on seeking the maximum sensitivity at a fixed specificity, it is a straightforward extension to obtain the maximum specificity at a given sensitivity. By assuming multivariate normal distributions for the diagnostic tests in both the diseased and healthy populations, an optimum linear combination test is searched numerically over all possible linear combinations under a binormal ROC setting. The method is exemplified with a real-world data on the diagnosis of pancreatic cancer. The performance of the method is also assessed with simulation studies. Results show that the proposed method can provide an accurate point estimate of the expected maximum sensitivity even in a relatively small sample size. The performance of the estimated confidence interval is also evaluated in terms of attaining the nominal 95% coverage based on the empirical coverage probability in the simulation study. The results show that a better coverage can be produced with moderate to large sample sizes.

The means and covariance for most populations are unknown in practice, and the corresponding maximum likelihood estimates (MLE) from the observed samples are frequently used. It is important to point out that the results from this work depend on the assumption of multivariate normality for the multiple diagnostic tests. In addition, our maximization process is based on the MLEs of the first two moments rather than individual measurements, and the proposed method may be relatively more sensitive to the normality assumption. In cases where a real-world data does not satisfy this assumption, some transformations may be necessary to improve normality and the proposed method can then be applied to the transformed data. When the normality assumption of X and Y fails, there will be in general some degeneration in the performance of our method, similar to that of a classical binormal ROC curve modeling. Note that the weights are dimensionless and thus are more appropriate for diagnostic tests with similar units. Otherwise certain data preparation (such as data transformation or normalization) is needed to reduce the dissimilarity among values from different tests. It also should be pointed out that, as explained by Anderson and Bahadur (1962), the method to identify optimum linear combination at a fixed specificity (sensitivity) may become problematic when the specificity (sensitivity) is extremely large. In such a case, we will work at a location of normal distribution that is far away from its central and thus the variance will dominate the estimation procedure.

References

Anderson, T. W. and Bahadur, R. R. (1962). Classification into two multivariate normal distributions with different covariance matrices. *The Annals of Mathematical Statistics* **33**, 420-431.

Begg, C. B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980s. *Statistics in Medicine* **10**, 1887-1895.

Cai, T, and Moskowitz, C. S. (2004). Semi-parametric estimation of the binormal ROC curve for a continuous diagnostic test. Biostatistics 4, 573-586.

DeLong ER, DeLong DM, and Clarke-Pearson DL (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* **44**, 837-45.

DeLong, E. R, Vernon, W. B. and Bollinger, R. R. (1985). Sensitivity and specificity of a monitoring test. *Biometrics* **41**, 947-58.

Greenhouse, S. W. and Mantel, N. (1950). Evaluation of diagnostic tests. *Biometrics* **6**, 399-412.

Gentle, J. E. (2002). *Elements of Computational Statistics.* Springer-Verlag.

Hanley, J. A. (1989). Receiver operating characteristic (ROC) methodology: the state of the art. *Clinical Reviews in Diagnostic Imaging* **29**, 307-335.

Hanley, J. A. (1996). The use of "binormal" model for parametric ROC analysis of quantitative diagnostic tests. *Statistics in Medicine* **15**, 1575-1585.

Hanley JA, and McNeil BJ (1984). Statistical approaches to the analysis of ROC curves. *Medical Decision Making* **4**, 137-150.

Linnet, K. (1987). Comparison of quantitative diagnostic tests: type I error, power and sample size. *Statistical in Medicine* **6**, 147-158.

Ma, G, and Hall, W. J. (1993). Confidence bands for receiver operating characteristic curves. *Medical Decision Making* **13**, 191-197.

McClish, D. K. (1987). Comparing the areas under more than two independent ROC curves. *Medical Decision Making* **7**, 149-155.

Metz, C. E., Wang, P.-L, and Kronman, H. B. (1984). A new approach for testing the significance of differences between ROC curves measured from correlated data. In *Information Processing Medical Imaging VIII* (Edited by F. Deconick). The Hague.

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

Su, J. Q. and Liu, J. S. (1993). Linear combinations of multiple diagnostic markers. *Journal of the American Statistical Association* **88**, 1350-1355

Swets, J. A. and Pickett, R. M. (1982). *Evaluation of Diagnostic Systems: Methods from Signal Detection Theory.* Academic Press.

Venkatraman, E. S. and Begg, C. B. (1996). A distribution-free procedure for comparing receiver operating characteristic curves from a paired experiment. *Biometrika* **83**, 835-848.

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

Xiong, C, McKeel, D. W., Miller, J. P. and Morris, J. C. (2004). Combining correlated diagnostic tests: application to neuropathologic diagnosis of Alzheimer's disease. *Medical Decision Making* **24**, 659-669.

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

Received June 2, 2006; accepted October 3, 2007.

Feng Gao
Division of Biostatistics
Washington University School of Medicine
660 S. Euclid Ave.
St. Louis, MO 63110, USA
feng@wustl.edu

Chengjie Xiong
Division of Biostatistics
Washington University School of Medicine
660 S. Euclid Ave.
St. Louis, MO 63110, USA
chengjie@wubios.wustl.edu

Yan Yan
Department of Surgery
Washington University School of Medicine
660 S. Euclid Ave.
St. Louis, MO 63110, USA
yan@wubios.wustl.edu

Kai Yu
Division of Cancer Epidemiology and Genetics
National Cancer Institute
Bethesda, MD 20892, USA
yuka@mail.nih.gov

Zhengjun Zhang
Department of Statistics
1221 Medical Sciences Center
1300 University Ave
Madison, WI 53706, U zjz@stat.wisc.edu