6-27-2011

When Does Combining Markers Improve Classification Performance and What Are Implications for Practice?

Aasthaa Bansal
University of Washington, abansal@uw.edu

Margaret Sullivan Pepe
Fred Hutchinson Cancer Rsrch Center, mspepe@uw.edu

1 Background

Biomarkers and clinical predictors are sought to assist in disease screening, in diagnosis and in making prognostic assessments for patients after diagnosis. This is an active area of research that has met with mixed success. In cancer research in particular, many biomarkers have been discovered but none, on its own, has yet been shown to have adequate performance for use in clinical practice. Efforts are currently underway to assemble panels of markers with the goal of developing marker combinations that have better performance.

We have previously investigated statistical characteristics of a single biomarker that lead to accurate classification or prediction of outcomes for individuals (Pepe et al., 2004). We and others have shown that the marker must be very strongly associated with outcome, having an odds ratio much larger than is typically observed in epidemiologic studies of association. This observation has been useful in setting targets and expectations for the performance of single biomarkers. In this paper, we turn our attention to the setting where the performance of the best-performing marker is still inadequate, and we seek to combine other markers with it. We ask what statistical characteristics of an additional marker lead to substantially improved performance. This might help develop strategies for assembling panels of markers for evaluation in rigorous validation studies.

As an example we consider CA-125, which is a biomarker for ovarian cancer. For early detection of ovarian cancer, CA-125 was recently shown to have the best performance among a panel of 6 markers studied (Anderson et al., 2010). Yet its performance is far from adequate for population screening. The area under the receiver operating characteristic (ROC) curve (AUC) is approximately 0.7 for detecting ovarian cancer 2 years prior to symptomatic clinical diagnosis. Moreover, using a threshold that sets the false positive rate (FPR) at 0.05, approximately 18% of cancers can be detected early with CA-125, i.e. the true positive rate (TPR) is only 0.18. Can we expect that combining other markers with CA-125 will improve performance substantially for screening? How should we identify markers for combining with CA-125?

A common intuitively appealing strategy for selecting novel markers to combine with each other and with existing markers is to identify those with good performance on their own. One

may prioritize those with low correlation since they apparently provide independent information. Moreover, in practice linear marker combinations are often studied. For example, Gail (2008) examined the potential improvement of the performance of standard clinical factors for predicting breast cancer that could be gained by adding seven SNPs with good performance on their own, assuming statistical independence with standard risk factors and using a linear logistic model for combination. He found that the area under the ROC curve increased only slightly from 0.607 to 0.632. As another example, Anderson et al. (2010) recently studied a panel of 6 ovarian cancer markers. Markers were selected for inclusion in this panel if they had high ranking performance on their own and combinations were studied primarily using linear algorithms.

In the first part of this paper we quantify the potential gain in classification performance that can be achieved by combining markers assuming various classic statistical and biologically motivated models for their joint distributions. We also note the function for combining them optimally and compare with the linear combination function. Practical implications of our results are discussed in the second part of this paper. In particular, we show that a broader approach to selecting panels of markers than the current popular approach may lead to better marker combinations but will require large sample sizes in order to be fruitful.

2 Marker Distributions and Combinations

2.1 Methods

We study marker combinations for classifying subjects according to true outcome status D, with $D = 1$ denoting a case and $D = 0$ a control. We use a variety of statistical models for the joint distributions of markers and quantify what can possibly be achieved by combining them. Our calculations are done for the models themselves, not for estimates based on a sample of data. Sampling variability in real data adds another layer of complexity that we address later. For simplicity and to gain insights we focus on combining two markers. One of the markers, that we call the standard or baseline marker X, has specified performance on its own. We consider another marker that we call the novel marker Y, vary the performance of Y relative to X, and determine to what

extent the performance of the combination (X, Y) is improved relative to X alone. The sensitivity (TPR) and specificity (1-FPR) and the ROC curve, which is a plot of TPR versus FPR, are used to quantify classification performance.

Long established statistical theory (McIntosh and Pepe, 2002; Neyman and Pearson, 1933) has shown that the optimal combination of (X, Y) is obtained by calculating the risk score function of X and Y, defined as $r(X, Y) = P(D = 1|X, Y)$, and setting criteria for positivity as

$$r(X, Y) \; > \; threshold.$$

In other words, the ROC curve for $r(X, Y)$ is higher and further to the left than the ROC curve for any other combination of (X, Y). Therefore, we calculate the optimal combination function, $r(X, Y)$, for each model and compare ROC values for $r(X, Y)$ with corresponding ROC values for X in order to quantify the improvement in performance gained by combining Y with X.

2.2 Results

2.2.1 Bivariate Binormal Equal Correlation Markers

The bivariate binormal model is a classic statistical model for two diagnostic tests or markers (Metz and Kronman, 1980). It assumes that the markers, (X, Y), have a bivariate normal distribution in cases and controls. Although we focus on normal models, note that the results apply much more generally to markers that are normal after some unspecified monotone transformations, since the ROC curve is unchanged by such transformations of the data (Pepe, 2003). We further assume here that case and control distributions have the same covariance matrix. Without losing generality, in controls the markers have means 0 and standard deviations 1 since we can always standardize the markers using the control distribution. We write

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \text{ in controls}$$

and

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} \mu_X \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right) \text{ in cases.}$$

The ROC curve for X alone is binormal with intercept μ_X and slope 1

$$\text{ROC}_X(f) = \Phi\{\mu_X + \Phi^{-1}(f)\}$$

because $X \sim N(0,1)$ in controls and $X \sim N(\mu_X, 1)$ in cases (Pepe, 2003). The AUC for X is

$$\text{AUC}_X = \Phi(\mu_X/\sqrt{2}).$$

With μ_X = 0.742, we have AUC_X = 0.7 and $\text{ROC}_X(0.05)$ = 0.183, implying that using a threshold that corresponds to FPR = 5%, the test detects 18.3% of the cases. This reflects the approximate performance of CA-125 for detecting ovarian cancer 2 years before clinical symptoms.

The ROC curve and AUC for Y alone take the same forms as those for X, but with μ_Y replacing μ_X. Values of μ_Y in the range 0 to 0.742 were considered as they correspond to performance ranging between that of a useless marker (AUC = 0.5; ROC(0.05) = 0.05) and that of the standard marker X (AUC = 0.7; ROC(0.05) = 0.183).

For the bivariate binormal model with equal correlation, it is well known that

$$r(X,Y) = \text{expit}(\alpha_0 + \alpha_1 X + \alpha_2 Y)$$

where $\alpha_1 = (\mu_X - \rho\mu_Y)/(1-\rho^2)$, $\alpha_2 = (\mu_Y - \rho\mu_X)/(1-\rho^2)$, and $\text{expit}(W) = \exp(W)/\{1+\exp(W)\}$. The combination of X and Y with optimal ROC curve is therefore $W = \alpha_1 X + \alpha_2 Y$, noting that the ROC curve for $\text{expit}(W)$ is the same as that for W because expit is a monotone increasing function. Since a linear combination of normal variables is normally distributed, it follows that the best combination has a binormal ROC curve

$$\text{ROC}_{X,Y}(f) = \Phi\left\{\frac{\mu_X + \frac{\alpha_2}{\alpha_1}\mu_Y}{\sqrt{1 + \frac{\alpha_2^2}{\alpha_1^2} + \frac{2\alpha_2\rho}{\alpha_1}}} + \Phi^{-1}(f)\right\}$$

and

$$\text{AUC}_{X,Y}(f) = \Phi\left\{\frac{\mu_X + \frac{\alpha_2}{\alpha_1}\mu_Y}{\sqrt{2\left(1 + \frac{\alpha_2^2}{\alpha_1^2} + \frac{2\alpha_2\rho}{\alpha_1}\right)}}\right\}.$$

When $\rho = 0$, the expressions simplify to

$$\text{ROC}_{X,Y}(f) = \Phi\left\{\sqrt{{\mu_X}^2 + {\mu_Y}^2} + \Phi^{-1}(f)\right\}$$

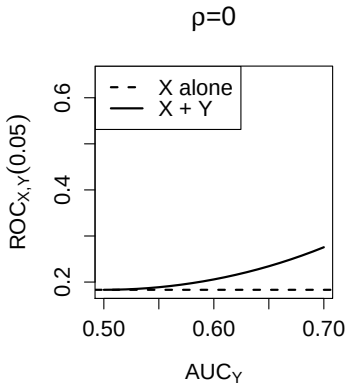
and

$$\text{AUC}_{X,Y} = \Phi\left(\sqrt{\frac{{\mu_X}^2 + {\mu_Y}^2}{2}}\right).$$

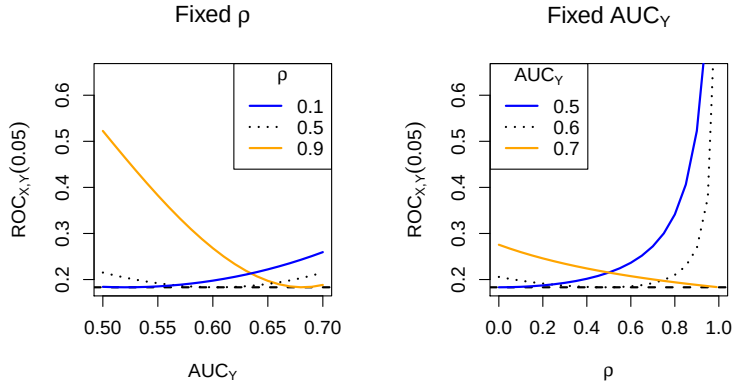
Therefore, when ρ = 0, larger values of μ_Y lead to higher ROC values for the combination. In other words, if a marker is uncorrelated with X, its marginal performance determines the performance of the combination — markers that perform better on their own lead to better performing combinations. This is not surprising. However, the magnitude of improvement in performance is surprisingly small (Figure 1(a)). Fixing the FPR at 0.05, we see that when the AUC for Y is 0.6, only 2.3% more cases are detected, i.e. $\text{ROC}_{X,Y}(0.05)$ = 0.206 versus $\text{ROC}_X(0.05)$ = 0.183. When Y on its own performs as well as X, (AUC_Y = 0.7), only 9.3% more cases are detected, i.e. $\text{ROC}_{X,Y}(0.05)$ = 0.276 versus $\text{ROC}_X(0.05)$ = 0.183. For AUC_Y = 0.6, Figure 2(a) left panel illustrates marker values classified as positive or negative using the optimal rule. See Supplementary Figure S1 for plots analogous to those in Figure 1, pertaining to $\text{AUC}_{X,Y}$ rather than $\text{ROC}_{X,Y}(0.05)$ as the measure of combination performance, though we consider $\text{AUC}_{X,Y}$ less clinically relevant.

Figure 1(b) concerns markers that may be correlated with X, $\rho > 0$. Interestingly, a positive correlation leads to a U-shaped curve for performance of the combination as a function of the marginal performance of Y (left panel). The inflection point occurs at $\text{AUC}_Y = \Phi(\rho\mu_X/\sqrt{2})$, i.e. $\mu_Y = \rho\mu_X$, where the coefficient α_2 for Y is equal to 0 and the optimal combination is determined to be X alone. From the right panel of Figure 1(b), we see that for markers with performance as good as X, i.e. AUC_Y = 0.7, performance improvement is maximized when ρ = 0 and decreases to no improvement when ρ = 1. In the latter setting, Y provides the same information as X so it adds nothing over X alone. Interestingly, however, for markers that are poor performers on their own ($\text{AUC}_Y < 0.7$), the existence of a correlation with X in cases and in controls yields a combination with improved performance. In the extreme, when AUC_Y = 0.5 and ρ = 1, the combination is a

(a) No Correlation



(b) Equal Correlation



(c) Unequal Correlation

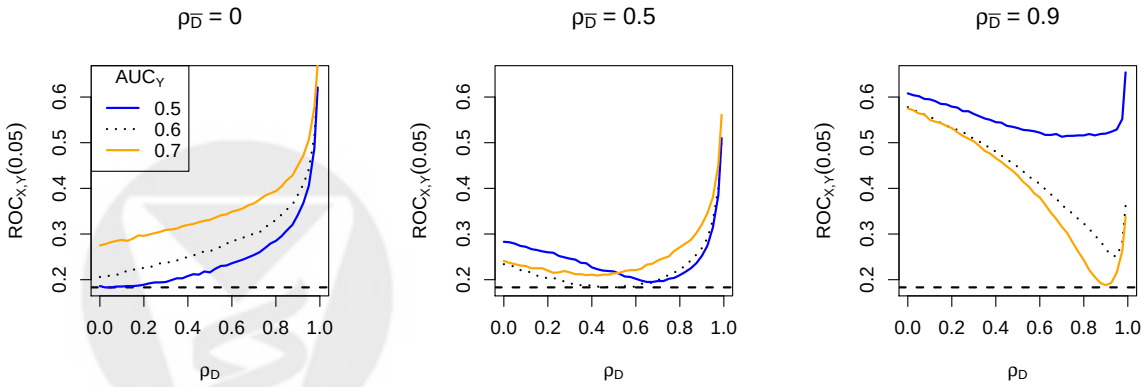
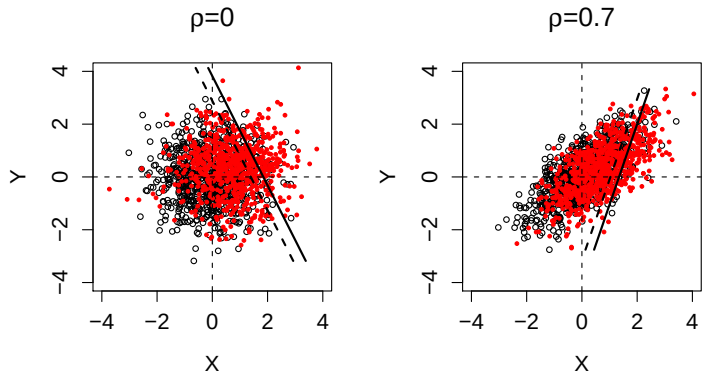
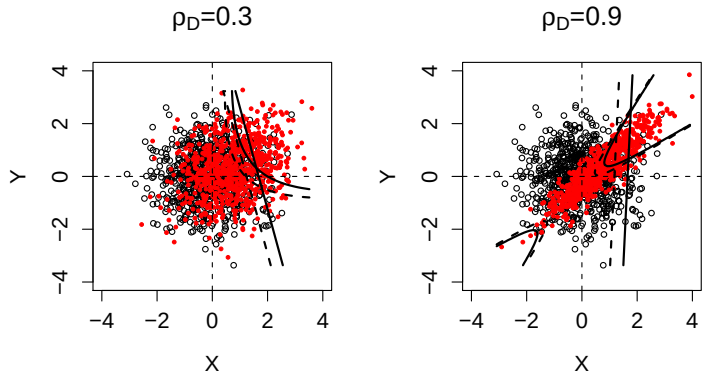


Figure 1: Bivariate Binormal Markers - Detection of cases by the combination (X,Y) at the threshold that leads to FPR = 0.05. The baseline marker X alone detects 18% of cases (AUC$_X$ = 0.7) and is indicated by a black dashed line. Shown are settings where (a) the correlation between markers in both cases and controls is 0, (b) the correlation is equal in cases and controls with positive coefficient ρ, and (c) the markers have a different correlation in controls and in cases, with correlation coefficients $\rho_{\bar{D}}$ and ρ_D, respectively. See Supplementary Figure S1 for analogous plots pertaining to AUC$_{X,Y}$ as the measure of combination performance.

(a) Equal Correlation



(b) Unequal Correlation ($\rho_{\bar{D}} = 0$)



(c) Mixture Model

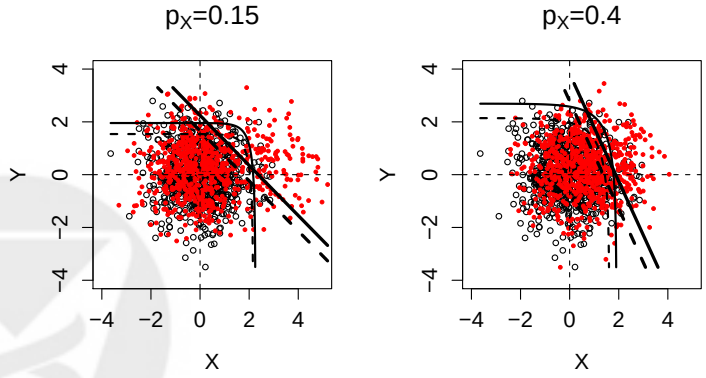


Figure 2: Decision boundaries that separate positive and negative classifications based on (X,Y) when $\text{ROC}_Y(0.05) = 0.100$ (or equivalently when $\text{AUC}_{X,Y} = 0.6$ for (a) and (b)). The FPR = 0.05 and FPR = 0.10 boundaries are shown with solid and dashed curves, respectively. Both the optimal and linear boundaries are shown. The solid points represent cases, while the hollow circles represent controls. Shown are settings where (a) the correlation between markers is equal in cases and controls, (b) the markers have a different correlation in controls and in cases, with correlation coefficients $\rho_{\bar{D}}$ (set to 0) and ρ_D, respectively, and (c) the markers have a mixture bivariate binormal distribution in cases, with X and Y being discriminatory in proportions p_X and p_Y of cases, respectively. We set $p_Y = 0.4$ and $p_X = 0.15, 0.4$. See Supplementary Figure S2 for more examples.

perfect marker. The classification of marker values as positive or negative when AUC_Y = 0.6 and ρ = 0.7 is shown in Figure 2(a) right panel.

The fact that poorly performing but highly correlated markers can add substantially to marker performance is an intriguing observation that we found surprising initially. However, such phenomena can arise in several practical settings. Consider that a measurement X made from a biological sample may be comprised of two biologic components, $X = W+Y$. If the component W is an excellent marker that cannot be measured and Y is a component that is unrelated to disease, then Y is a useless marker on its own, but in combination with X it yields W, the excellent marker. For example, the biomarker prostate-specific antigen (PSA) is made up of free and bound PSA components. As another example, consider the setting where Y is a baseline measurement of a biomarker, X is the current value and W is the change in the marker from baseline to present. If the change in the marker is strongly associated with occurrence of disease, the current measure X along with a possibly uninformative baseline value Y yields the change, W, which is an excellent marker. As examples, it has been hypothesized that the change in CA-125 and in PSA may be more informative of ovarian cancer and prostate cancer, respectively, than values measured at one point in time (Berger et al., 2005; McIntosh, Urban, and Karlan, 2002; Skates et al., 1995; Skates and Pauler, 2001; Slate and Cronin, 1997).

2.2.2 Bivariate Binormal Unequal Correlation Markers

Two markers may have a different correlation in cases than in controls. For example, there may be a high correlation in cases but a low correlation in controls. This would occur if both markers are at higher levels in cases with more extensive disease and at lower levels in cases with less extensive disease, a likely scenario for many markers. Markers of cell growth and inflammation fit this sort of scenario. For this setting we consider the following general joint model:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{\bar{D}} \\ \rho_{\bar{D}} & 1 \end{pmatrix}\right) \text{ in controls}$$

and

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{D} \\ \rho_{D} & 1 \end{pmatrix}\right) \text{ in cases,}$$

where subscripts $\bar{D}$ and D are used for controls and for cases, respectively. The optimal marker combination is no longer linear when $\rho_D \neq \rho_{\bar{D}}$. The risk function $r(X, Y)$ can be shown to be a monotone function of $(\frac{1}{1-\rho_D^2} - \frac{1}{1-\rho_{\bar{D}}^2})X^2 + (\frac{1}{1-\rho_D^2} - \frac{1}{1-\rho_{\bar{D}}^2})Y^2 - 2(\frac{\rho_D}{1-\rho_D^2} - \frac{\rho_{\bar{D}}}{1-\rho_{\bar{D}}^2})XY +$
$\frac{1}{1-\rho_D^2}\{-2(\mu_X - \rho_D\mu_Y)X - 2(\mu_Y - \rho_D\mu_X)Y + (\mu_X^2 + \mu_Y^2 - 2\rho_D\mu_X\mu_Y)\}$.

Figure 2(b) shows marker values classified as positive and negative using this non-linear optimal combination for different values of ρ_D, assuming $\rho_{\bar{D}} = 0$. We see that discordance of marker values leads to negative classification while concordance leads to positive classification for disease. That is, if one marker is high but is not confirmed by a high value for the other marker, the subject is unlikely to be classified as a case. The result makes sense intuitively for this model.

Figure 1(c) shows the increment in performance gained by combining Y with X when the correlation is unequal between cases and controls. These calculations were made using large simulations since an analytic formula for $\text{ROC}_{X,Y}(f)$ was not feasible in this setting. In the left panel corresponding to the special setting where $\rho_{\bar{D}} = 0$ and $\rho_D > 0$, we see that better marginal performance of Y always leads to better performance for the combination (X, Y) and that stronger correlation of the markers in cases leads to better performance. Much larger gains in performance are possible when Y is correlated with X in cases ($\rho_D > 0$) compared with when $\rho_D = 0$, the setting shown in Figure 1(a). For example, if $\text{AUC}_Y = 0.7$ and it is uncorrelated with X in cases (Figure 1(a)), the combination can detect only 9.3% more cases than X alone ($\text{ROC}_{X,Y}(0.05) = 0.276$ versus $\text{ROC}_X(0.05) = 0.183$). However, from the left panel of Figure 1(c) we see that it can detect 17% more cases ($\text{ROC}_{X,Y}(0.05) = 0.349$) when $\rho_D = 0.6$ and 22% ($\text{ROC}_{X,Y}(0.05) = 0.400$) when $\rho_D = 0.8$. Moreover, observe that even when Y is not useful on its own ($\text{AUC}_Y = 0.5$), it can greatly improve the performance of X if it is correlated with X only in cases.

When markers have a positive correlation in controls, we see from the right two panels of Figure 1(c) that the performance of (X, Y) combined is a U-shaped function of ρ_D. Better performance occurs when ρ_D and $\rho_{\bar{D}}$ are very different. This also leads to the somewhat unintuitive result that when markers are highly correlated in controls, $\rho_{\bar{D}} = 0.9$, worse marginal performance of Y leads to better performance of the combination.

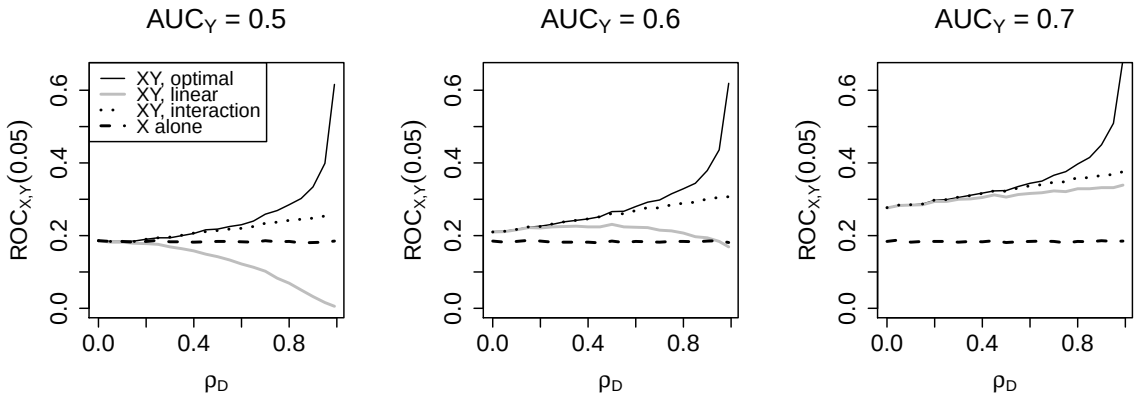


Figure 3: Bivariate Binormal Unequal Correlation Markers - Comparison of decision rules in detecting cases when FPR = 0.05. Shown here are (X,Y) combinations using the optimal risk score, a logistic model with linear terms and a logistic model with linear and interaction terms. The baseline marker X alone detects 18% of cases and is indicated by a black dashed line. Markers have 0 correlation in controls and correlation ρ_D in cases.

We noted earlier that the optimal marker combination is non-linear when marker correlations in cases and controls are unequal (Figure 2(b)). However, it is common practice to use the coefficients from a linear logistic model to obtain a combination of the form $\hat{\alpha_1}X + \hat{\alpha_2}Y$. Figure 2(b) shows marker values classified as positive and negative at FPR = 0.05 and FPR = 0.10, using such linear combinations derived from the logistic likelihood. We see that the structure imposed by unequal correlation in cases and controls implies that linear decision boundaries do not approximate very well the non-linear optimal boundaries. Figure 3 compares the combinations with respect to TPR when FPR is set equal to 0.05. We see from the right two panels that the linear combination has performance comparable with the optimal combination for AUC$_Y \geq 0.6$ when $\rho_D < 0.5$, but with $\rho_D > 0.5$ it has relatively poor discriminatory ability. Interestingly, when AUC$_Y$ = 0.5, the linear combination performs even worse than X alone. We also find that when an interaction term is added to the model, performance is often improved significantly, but is not comparable with the optimal combination when correlation between X and Y is high in cases but zero in controls.

2.2.3 Mixture Bivariate Binormal Markers

Diseases are often heterogeneous in nature. This is particularly true for cancer where it is thought that unknown subtypes exist. Correspondingly, different biomarkers may be associated with different subtypes. The following statistical model incorporates this concept:

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim \text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \text{ in controls}$$

and

$$\begin{pmatrix} X \\ Y \end{pmatrix} \sim p_X \text{BVN}\left(\begin{pmatrix} \mu_X \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) + p_Y \text{BVN}\left(\begin{pmatrix} 0 \\ \mu_Y \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right)$$
$$+(1 - p_X - p_Y)\text{BVN}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \right) \text{ in cases.}$$

That is, in cases the distribution for (X, Y) is a mixture of three distibutions. An interpretation is that X is discriminatory in a proportion p_X of cases, while Y is discriminatory in a different proportion p_Y of cases. The ROC curves for X and Y alone are mixture binormal. In particular, for X we have:

$$\text{ROC}_X(f) = p_X \Phi\{\mu_X + \Phi^{-1}(f)\} + (1 - p_X)f$$

with corresponding AUC:

$$\text{AUC}_X = p_X \Phi(\mu_X/\sqrt{2}) + (1 - p_X)/2.$$

The ROC and AUC for Y take the same forms, but with μ_Y replacing μ_X and p_Y replacing p_X. Observe that there are two parameters that define the marginal performance of each marker, (p_X, μ_X) for X and (p_Y, μ_Y) for Y. We set the performance of X as before, such that $\text{ROC}_X(0.05) = 0.183$, which is achieved under two configurations that we consider here, $(p_X, \mu_X) = (0.15, 3.17)$ and $(p_X, \mu_X) = (0.4, 1.35)$. The marginal performance of Y ranges from useless, $\text{ROC}_Y(0.05) = 0.05$, to that of X, $\text{ROC}_Y(0.05) = 0.183$. By fixing $p_Y = 0.15$ this corresponds to μ_Y ranging from 0 to 3.17 while setting $p_Y = 0.4$ this corresponds to μ_Y ranging from 0 to 1.35. Note that for a fixed overall performance criterion, smaller values for the proportion of marker-specific cases, p, correspond to larger values of μ, which in turn correspond to better detection of those marker-specific

cases. Moreover, note that the proportion of marker-specific cases imposes a bound on the maximum performance that can achieved by that marker. For example, when $p_Y = 0.15$, the maximum $\text{ROC}_Y(0.05)$ is 0.193, no matter how large μ_Y may be. Intuitively, this makes sense in the given context of heterogeneous diseases, where a marker may have good discriminatory ability within a particular subtype, but if that subtype occurs very rarely the marker may not prove to be very useful in the overall classification of subjects as positive or negative for the disease.

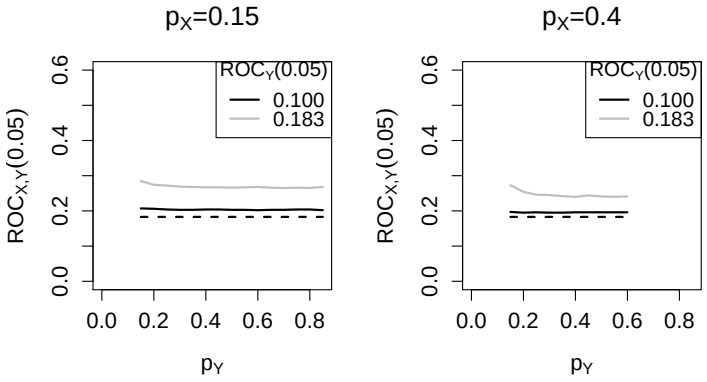
Figure 4 shows the increment in performance gained by combining Y with X. Again, these calculations were made using large simulations since deriving an analytic formula for $\text{ROC}_{X,Y}(f)$ was not feasible. Gains in performance are small when Y is comparable with X ($\text{ROC}_Y(0.05) = 0.183$), about 10% when $p_X = 0.15$ and 8% when $p_X = 0.4$ (Figure 4(a)). The gains are miniscule when $\text{ROC}_Y(0.05) = 0.100$. It is interesting that for $\text{ROC}_Y(0.05) = 0.183$, we see a somewhat larger performance increment when $p_X = 0.15$ ($\text{ROC}_{X,Y}(0.05) = 0.274$ at $p_Y = 0.25$), compared to when $p_X = 0.4$ ($\text{ROC}_{X,Y}(0.05) = 0.249$ at $p_Y = 0.25$). Moreover, while $\text{ROC}_{X,Y}(0.05)$ generally stays constant over varying p_Y, there is a slight upward spike observed for $p_Y < 0.2$. This result is noteworthy since it implies that for a fixed marginal performance, markers with lower subtype prevalence are to be preferred. That is, the larger separation of marker values between marker-specific cases and controls for $p_Y = 0.15$ makes a joint classification rule that is more effective. This idea is illustrated in Figure 4(b), where a larger increment occurs for $p_Y = 0.15$ than for $p_Y = 0.4$. An analogous result holds for p_X and we see that the largest increment of 10% occurs for the setting where both subtypes have lower prevalence but have markers that are highly sensitive. Nevertheless we see that the increment in performance is fairly small in all settings for the mixture bivariate binormal setting.

As shown in Figure 2(c), the optimal marker combination in the mixture bivariate binormal setting is again non-linear. The risk function, r(X,Y), is a monotone function of

$$p_X \text{exp}(-\mu_X^2/2 + \mu_X X) + p_Y \text{exp}(-\mu_Y^2/2 + \mu_Y Y)$$

In Figure 2(c), we use this optimal combination as well as a linear combination based on a logistic model to classify marker values as positive or negative at FPR = 0.05 and FPR = 0.10. While the

(a) Fixed $\text{ROC}_Y(0.05)$



(b) Fixed p_Y

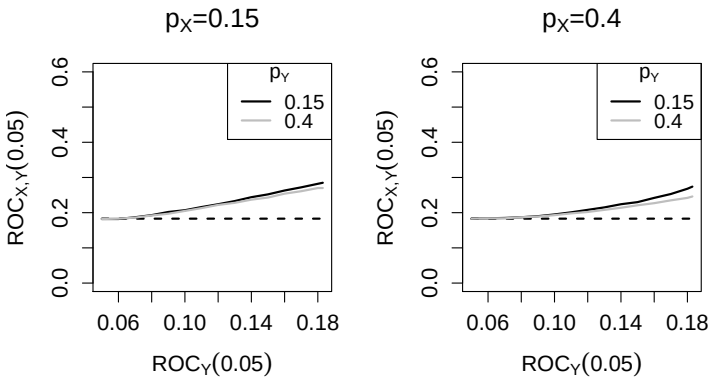


Figure 4: Mixture Binormal Markers - Detection of cases by the combination (X,Y) and the threshold that leads to FPR = 0.05. The performance of X alone is indicated by a black dashed line that remains constant at $\text{ROC}_X(0.05) = 0.183$. Shown here are results for when (a) $\text{ROC}_Y(0.05)$ is fixed and the combination performance is observed over p_Y varying from 0.15 to 1 - p_X and (b) p_Y is fixed and $\text{ROC}_Y(0.05)$ is varied from 0 to 0.183.

flexibility of the optimal rule fits the structure of the data better, a linear boundary can perform relatively well here (also see Supplementary Figure S3). When $p_X = 0.15$ and $\text{ROC}_Y(0.05) = 18.3\%$, $\text{ROC}_{X,Y}(0.05) = 29\%$ using the optimal rule versus $\text{ROC}_{X,Y}(0.05) = 24\%$ using the linear rule. The linear rule gains about half the increment gained by the optimal rule across all values of p_Y. When $p_X = 0.4$ and $\text{ROC}_Y(0.05) = 18.3\%$, $\text{ROC}_{X,Y}(0.05) = 27\%$ using the optimal rule versus $\text{ROC}_{X,Y}(0.05) = 23\%$ using the linear rule in the scenario where $p_Y = 0.15$. As p_Y increases, the performance of the linear rule closely approximates that of the optimal rule.

2.3 General Implications of Numerical Studies

In the above numerical studies we investigated the incremental value of Y under three specific scenarios for the joint behaviors of X and Y. We found that under the mixture bivariate binormal model and the classic bivariate binormal model with equal variances in cases and controls and zero correlation, the increment in performance is typically very small. This holds true even when Y's performance is comparable with that of X. We found much larger gains in performance under alternative bivariate binormal scenarios including: novel markers with poor marginal performance but highly correlated with X conditional on disease status; and novel markers uncorrelated with X in controls but highly correlated with X in cases (and vice versa).

Our simulations covered a very limited set of scenarios. In practice, markers may not follow the mixture model or the binormal model or they may have different variances in cases versus controls for example. Therefore, our results do not provide specific guidance on which markers are likely to yield substantial improvements in performance in practice. Rather they suggest casting a wide net in the search for novel markers. Restricting the search for novel markers to those that have good marginal performance or low correlation, as is common practice, is ill advised. We also showed that combining markers with linear rules may be too restrictive. Much better combinations than linear combinations existed in several of the scenarios we studied.

3 Comparing Broad and Standard Strategies for Selecting Marker Panels

We now turn to strategies for using data to select a small subset of markers from a large set of markers for further study with the goal of developing a marker combination with good classification performance. The standard approach is to rank markers according to their marginal classification performance, to select the top K ranking markers, and to evaluate all combinations of those K markers. The results of our numerical studies suggest that this approach may miss markers that perform extremely well in combination. Therefore we investigate here a broader selection

strategy that evaluates combinations of all available markers regardless of marginal marker performance. To keep the problem simple and hopefully more insightful, we consider only pairwise combinations as above.

3.1 Colon Cancer Dataset

Autoantibodies to tumor antigens may be useful for cancer screening. We use a dataset comprised of 2100 protein microarray autoantibody measurements from 70 case subjects with colon cancer and 70 control subjects. We transformed each marker so that its distribution in controls was normal with mean 0 and standard deviation 1. We used a random set of 100 of the 2100 markers measured on a random subset of 70 subjects. The standard and broad strategies were applied. These 70 subjects were used as the training dataset to select the top P = 10 marker combinations under both strategies. We reserved the data on the remaining 70 subjects to act as a test dataset for performance evaluation.

In the standard strategy, all pairwise combinations of the K = 10 markers with best marginal performances were considered. Pairs were combined using logistic regression and their AUCs estimated with cross validation. The P = 10 combinations with highest cross-validated AUCs were then evaluated on the independent test dataset in order to determine the performance of combinations derived from the standard strategy. In the broad strategy, logistic regression was applied to all pairs regardless of their marginal performance and the P = 10 combinations with highest cross-validated AUCs were evaluated in the test dataset. We used logistic regression with interaction terms in one analysis and logistic regression without interaction terms in another in order to determine if a more complex model yielded better combinations.

Results for a single dataset (see Supplementary Table S1) show that the broad strategy yields worse combinations than the standard strategy in this setting. However, no general conclusions can be drawn from results of one dataset as we found substantial variation in results with the split of the dataset into training and test components and with the random sample of 100 markers included. Therefore we repeated this exercise 100 times and summarized the results in Figure 5(a). Each

Table 1: Proportion of repetitions in which the kth ranking combination in the training dataset had a higher test dataset AUC from the first strategy than from the second strategy.

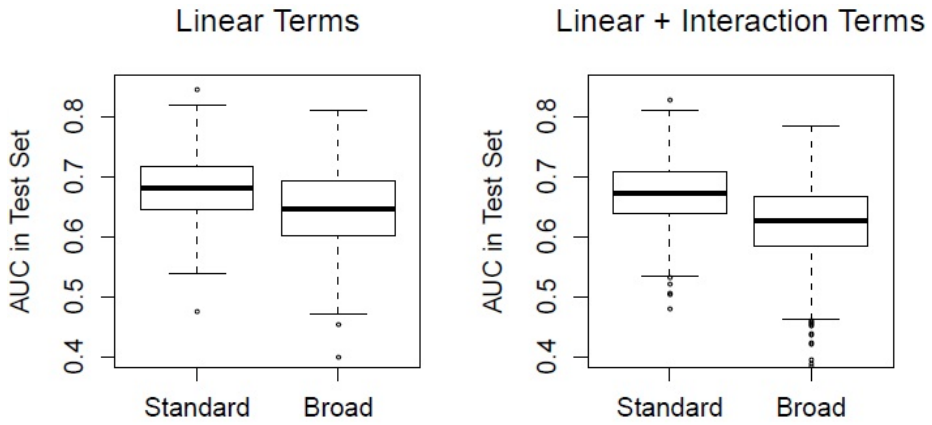
(a) Original Colon Cancer Dataset (N = 140)

Marker Rank	Broad, no interaction > Standard, no interaction	Broad, w/ interaction > Standard, w/ interaction	Broad, w/ interaction > Broad, no interaction
1	0.22	0.28	0.44
2	0.29	0.21	0.38
3	0.29	0.32	0.38
4	0.33	0.26	0.35
5	0.32	0.31	0.35
6	0.35	0.24	0.37
7	0.27	0.24	0.37
8	0.32	0.26	0.35
9	0.28	0.26	0.45
10	0.36	0.26	0.37

(b) Expanded Colon Cancer Dataset (N = 7,000)

Marker Rank	Broad, no interaction > Standard, no interaction	Broad, w/ interaction > Standard, w/ interaction	Broad, w/ interaction > Broad, no interaction
1	0.89	0.95	0.74
2	0.95	0.98	0.75
3	0.89	0.91	0.67
4	0.89	0.98	0.74
5	0.96	0.96	0.73
6	0.90	0.92	0.77
7	0.89	0.91	0.74
8	0.89	0.94	0.69
9	0.84	0.90	0.71
10	0.93	0.94	0.71

(a) Original Colon Cancer Dataset (N = 70 in the training dataset)



(b) Expanded Colon Cancer Dataset (N = 3,500 in the training dataset)

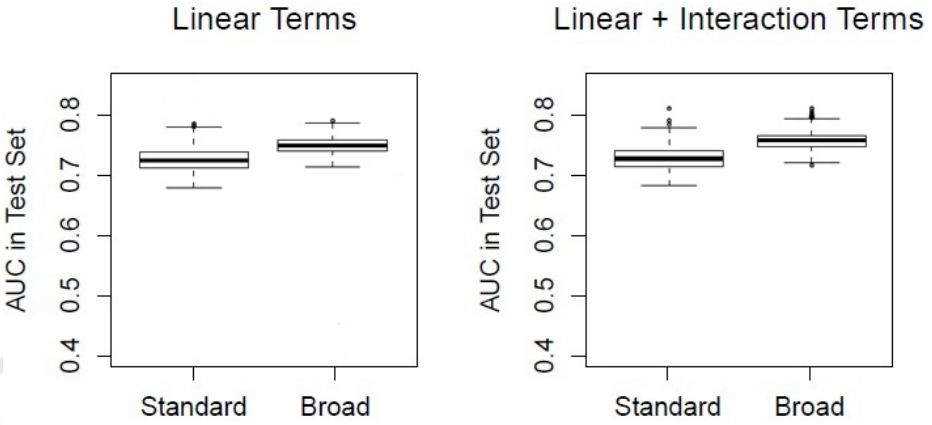


Figure 5: Comparison of standard and broad strategies, using 100 repetitions of different sets of 100 markers and training-test data from the (a) original and (b) expanded colon cancer datasets.

boxplot contains 1000 AUC values, showing distributions of the AUCs estimated with the test dataset for the top 10 marker combinations derived from the training dataset. Table 1(a) shows the proportion of repetitions in which the kth ranking combination derived from the broad strategy in the training dataset had a higher test dataset AUC than that for the kth ranking combination derived from the standard strategy. We see that in general the broad strategy yields worse marker combinations than the standard strategy, not better combinations as we had hoped. Use of more complex combinations that include an interaction term led to even poorer performance. These disappointing results however were reversed when we repeated the exercise in a larger dataset. Since we did not have access to a large real dataset, we expanded the colon cancer dataset by a factor of 50 by replicating it and adding to each biomarker value normally distributed noise with mean 0 and standard deviation 0.01. Results shown in Figure 5(b) and Table 1(b) show that in this much larger dataset, the broad strategy finds better performing marker combinations than does the standard strategy. For example, the top marker pair derived with the broad strategy had better performance than the top marker pair derived with the standard strategy in 95% of the repetitions when a logistic model with interaction was used for combining markers. Moreover, including an interaction term in the logistic model was beneficial in this dataset. We see that using the broad strategy, the top marker combination derived from the model including an interaction had AUC higher than the top marker combination derived from the model without interaction in 74% of the repetitions.

4 Discussion

Results from the colon cancer dataset suggest that a broad search strategy can be useful in identifying markers for combination. More dramatic results were observed with simulated datasets where markers with joint distributions discussed earlier in this paper were included among candidate markers available (See Appendix A). Those results also showed advantages for including interaction terms in the combination score. Advantages of the broad search and flexible score are however only realized when large datasets are available for identifying the combinations. In

current practice, datasets for identifying markers are usually small. We propose that much larger sample sizes be used for these studies in the future.

How should one choose the sample size of a marker panel identification study? One approach is to perform simulation studies based on hypothesized joint distributions for biomarkers. By varying the size of the simulated dataset one can determine the sample size at which good marker combinations are likely to rank highly and are therefore likely to be pursued in further studies. If pilot data are available one could base simulation studies on that. For example, suppose the colon cancer dataset were pilot data and we were to design another panel identification study, we could expand the pilot dataset as we did earlier, and determine at what expanded size the selected combination or combinations are highly likely to have performance that is at or above some target.

Note that in practice combining markers does not always lead to improvements in performance. In particular, non-optimal combinations may have worse performance than either marker on its own. We saw this in the left panel of Figure 3. A simple example involving the equal correlation bivariate binormal model is reported in Appendix B. One must be cognizant that sampling variability in coefficient estimates for risk models is one factor that leads to suboptimal marker combinations.

In summary, we have shown that the practice of restricting attention to markers with good marginal performance has the potential to miss certain marker combinations that perform extremely well. Although we investigated combinations of only two markers, the implication is clearly true for combinations of more than two markers as well. We showed that by broadening the strategy for assembling marker panels to include markers with poorer marginal performance that appear to contribute substantially to combination performance, better marker combinations may be found. However, large sample sizes will be required for these marker panel identification studies in order for the broadened strategy to be fruitful.

Acknowledgements

This work was partially supported by the National Institutes of Health [CA129934, CA086368, GM054438]. The authors are grateful to Dr Samir Hanash of the Fred Hutchinson Cancer Research Center for allowing us to use the colon cancer dataset. *Conflict of Interest:* None declared.

References

Anderson, G.L., McIntosh, M., Wu, L., Barnett, M., Goodman, G., Thorpe, J.D., Bergan, L., Thornquist, M.D., Scholler, N., Kim, N., O'Briant, K., Drescher, C., and Urban, N. (2010). Assessing Lead Time of Selected Ovarian Cancer Biomarkers: A Nested Case-control Study. *Journal of the National Cancer Institute* **102,** 26–38.

Berger, A.P., Deibl, M., Steiner, H., Bektic, J., Pelzer, A., Spranger, R., Klocker, H., Bartsch, G., and Horninger, W. (2005). Longitudinal PSA changes in men with and without prostate cancer: Assessment of prostate cancer risk. *The Prostate* **64,** 240–245.

Gail, M.H. (2008). Discriminatory accuracy from single-nucleotide polymorphisms in models to predict breast cancer risk. *Journal of the National Cancer Institute* **100,** 1037–1041.

Ma, S. and Huang, J. (2007). Combining multiple markers for classification using ROC. Biometrics 63, 751757. McIntosh, M.W. and Pepe, M.S. (2002). Combining several screening tests: optimality of the risk score. Biometrics 58, 657664. McIntosh, M.W., Urban, N., and Karlan, B. (2002). Generating longitudinal screening algorithms using novel biomarkers for disease. *Cancer Epidemiology, Biomarkers & Prevention* **11,** 159–166.

Metz, C.E. and Kronman, H.B. (1980). Statistical significance tests for binormal ROC curves. *Journal of Mathematical Psychology* **22,** 218–243.

Neyman, J. and Pearson, E.S. (1933). On the problem of the most efficient tests of statistical hypotheses. *Philosophical Transactions of the Royal Society London A* **24,** 289–337

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

Pepe, M.S., Janes, H., Longton, G., Leisenring, W., and Newcomb, P. (2004). Limitations of the odds ratio in gauging the performance of a diagnostic, prognostic, or screening marker. *Americal Journal of Epidemiology* **159,** 882–890.

Pepe, M.S., Cai, T., and Longton, G. (2006). Combining predictors for classification using the area under the receiver operating characteristic curve. *Biometrics* **62,** 221–229.

Skates, S.J., Xu, F.J., Yu, Y.H., Sjovall, K., Einhorn, N., Chang, Y., Bast Jr., R., and Knapp, R. (1995). Toward an optimal algorithm for ovarian cancer screening with longitudinal tumor markers. *Cancer Supplement* **76(10),** 2004–2010.

Skates, C.J. and Pauler, D.K. (2001). Screening based on the risk of cancer calculation from Bayesian hierarchical change-point models of longitudinal markers. *Journal of the American Statistical Association* **96,** 429–439.

Slate, E.H. and Cronin, K.A. (1997). Changepoint modeling of longitudinal PSA as a biomarker for prostate cancer. *School of Operations Research and Industrial Engineering, Technical Report* **1189**.

Appendix A: Simulated Dataset

We evaluated our methods on data simulated from the models presented in Sections 2.2.1 and 2.2.2. To do so, we generated 50 pairs of baseline and novel markers that had joint bivariate binormal distributions with equal and unequal correlations in cases and controls, varying from 0 to 0.9. In each pair, the baseline marker was set to have AUC = 0.7, while the marginal performance of the second marker ranged from AUC = 0.5 to AUC = 0.7 (specifically, AUC = 0.5, 0.55, 0.6, 0.65, 0.7). Using all five of these AUC values and ρ = 0, 0.5, 0.95, we generated 15 pairs from the equal correlation model. Using the same five AUC values for the second marker, we also generated data from the unequal correlation model: five pairs with $\rho_{\bar{D}}$ = 0 and ρ_D = 0.7 and 30 pairs using all six combinations of $\rho_{\bar{D}}$ = (0.1, 0.5, 0.95) and ρ_D = (0.1, 0.5, 0.95), where $\rho_{\bar{D}} \neq \rho_D$. Along with 10 additional markers that were normally distributed noise (mean 0 and standard deviation 1) independent of all other markers, a total set of 110 markers was generated.

A fairly small training dataset of 250 case subjects and 250 control subjects (N = 500) was created in order to illustrate and compare the standard restrictive approach to selecting markers for combination with a broader strategy. To investigate the effect of using a larger dataset for marker combination selection, we repeated the analysis with a dataset of 500 case subjects and 500 control subjects (N = 1000). In both cases, we calculated the performance of selected combinations using Monte Carlo calculations with a very large simulated dataset. We used the area under the ROC curve here to quantify discrimination since it has less variability in small samples than estimated points on the ROC curve.

As discussed in the text for the colon cancer dataset, in the standard strategy, all pairwise combinations of the K = 10 markers with best marginal performances in the training dataset were considered. Pairs were combined using logistic regression in the training dataset and their AUCs estimated with cross validation. The P = 10 combinations with highest cross-validated AUCs in the training dataset were then evaluated on the independent large dataset in order to determine the performance of combinations derived from the standard strategy. In the broad strategy, logistic regression was applied to all pairs in the training dataset regardless of their marginal performance and

the P = 10 combinations with highest cross-validated AUCs were evaluated on the large dataset. We used logistic regression with interaction terms in one analysis and logistic regression without interaction terms in another in order to determine if a more complex model yielded better combinations.

Results for a single dataset are shown in Table A1. We focus first on the smaller dataset of 500 subjects. For both logistic models, all of the top combinations selected by the standard strategy involve pairs of markers that have good marginal performance (AUC = 0.7) and that are uncorrelated with each other in cases and in controls. These combinations have AUC values ranging from 0.768 to 0.771 — a moderate improvement over the baseline AUC of 0.7. The standard strategy misses three better combinations that are selected by the broad strategy with AUC values ranging from 0.817 to 0.944. These are marker pairs that are highly correlated in cases and controls (ρ = 0.95) and where the second markers perform poorly on their own (AUC = 0.5, 0.55, 0.6). Note that all of the combinations selected using this smaller dataset (by both strategies and using both logistic models) only involve marker pairs that have equal correlation in cases and controls and therefore have optimal combinations that involve only linear terms.

Because the simulation model is known, we know the optimal marker combinations. The AUCs of the 10 optimal combinations are also listed in Table A1. The top two combinations come from equal correlation marker pairs with a poorly-performing second marker, and are in fact the same as the two best marker pairs selected by the broad strategy using the smaller dataset (N = 500). However, the combination derived from the dataset is suboptimal because of sampling variability in the logistic regression coefficient estimates. The rest of the optimal marker combinations correspond to marker pairs with unequal correlation in cases (ρ_D = 0.95) and controls ($\rho_{\bar{D}}$ = 0.1) and varying marginal performance of the second marker (AUC = 0.5 to AUC = 0.7). These pairs have non-linear optimal combinations, and therefore one may expect some of them to be selected using the logistic model with interactions. However, it seems that the dataset is too small to correctly fit the more complicated model and to select combinations based on it. This problem was amplified when we investigated a third logistic model including quadratic terms. Using a sample size of 500, it performed poorly compared to the two simpler models (data not shown).

For each model, we also explore asymptotic behavior using a large dataset (N = 100,000) for the selection of top combinations. For the same marker pairs, the large dataset tends to produce better estimates of the logistic regression coefficients and therefore combinations with slightly higher AUC values than those from a dataset of size N = 500. It is not surprising then, that when we increase the size of our training dataset to N = 1000, we see an increase in AUC values compared to when N = 500. Using the logistic model with interaction, the broad strategy is now able to select a marker pair with unequal correlation in cases (ρ_D = 0.95) and controls ($\rho_{\bar{D}}$ = 0.1), where the optimal combination includes an interaction term. The AUC based on N = 1000 is 0.787. The corresponding large sample asymptotic AUC is 0.802, much smaller than the optimal AUC of 0.898. This inconsistency between the asymptotic AUC and the optimal AUC is no surprise given the theoretical results we saw earlier. Recall Figure 3, where we illustrated that a logistic model with an interaction term alone does not estimate the optimal combination well for high ρ_D when $\rho_{\bar{D}}$ = 0.

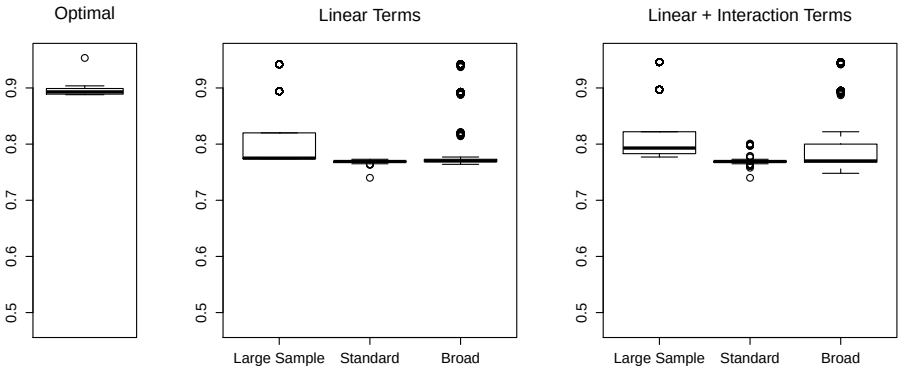
Results from 100 datasets are summarized in Figure A1, where for the standard and broad strategies, we show distributions of AUCs calculated using a large dataset for the top 10 marker combinations derived from the smaller datasets. In all, 1000 AUC values enter into the 'Standard' and 'Broad' box plots, 10 from each of the 100 repetitions. The boxplots labeled 'Large Sample' and 'Optimal' each contain AUC values for 10 marker combinations selected as described above, using a very large sample and the optimal risk score, respectively.

These results are consistent with the single dataset results. The broad strategy tends to select marker combinations that are at least as good as those selected by the standard strategy and both strategies perform slightly better when using a larger dataset (N = 1000 versus N = 500). The top combinations selected by the broad strategy tend to be of marker pairs that are highly and equally correlated in cases and controls with a second marker that has low marginal performance. The optimal combination in this case is linear and the simpler logistic model and smaller sample size (N = 500) seem to be adequate for selecting these combinations.

Based on the results for N = 1000, there appears to be little advantage of using a more complex

model over a model with linear terms only. The large sample performance, however, shows some improvement in AUCs with the more complicated model. This finding indicates that N = 1000 might not be large enough to reap the full benefits of including an interaction term. Moreover, as mentioned above, a model that includes only an interaction term is limited in how closely it approximates the optimal combination of markers with unequal correlation in cases and controls. Using an even larger sample size with quadratic terms would be expected to generate a larger increment in combination performance.

(a) N = 500 in the training dataset



(b) N = 1000 in the training dataset

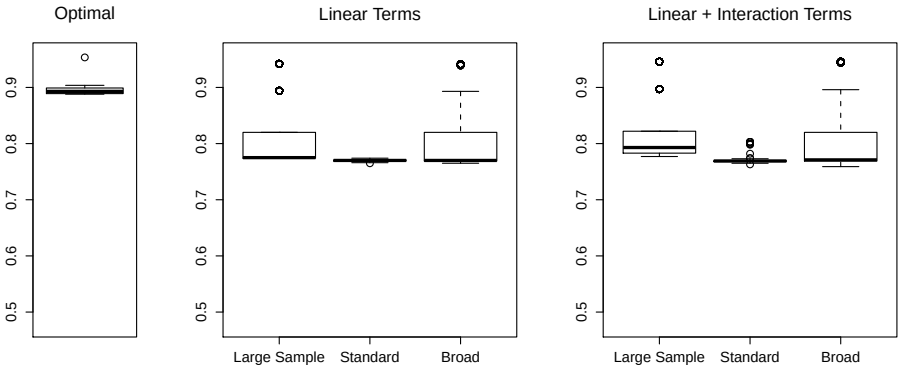


Figure A1: Results of 100 repetitions with 110 markers simulated based on the joint distributions presented in Sections 2.2.1 and 2.2.2. For logistic models with and without an interaction term, top ranking marker combinations were selected using the standard and broad strategies in training datasets of sizes (a) N = 500 and (b) N = 1000. For these combinations, AUC values calculated from a large dataset are presented here. The asymptotic behavior under each model is denoted by 'Large Sample' and shows top ranking combinations selected using a dataset of size N = 100,000. Also presented are top ranking combinations obtained using optimal risk scores. Marker combinations are different in different boxplots.

Table A1: Results of data analysis for a single dataset with 110 markers simulated based on the joint distributions presented in Sections 2.2.1 and 2.2.2. For logistic models with and without an interaction term, top ranking marker combinations were selected using the standard and broad strategies in training datasets of sizes N = 500 and N = 1000. For these combinations, AUC values calculated from a large dataset are presented here. The asymptotic behavior under each model is presented in the column labeled 'Large Sample', which shows top ranking combinations selected using a dataset of size N = 100,000. Also presented are top ranking combinations obtained using optimal risk scores. Marker combinations are different in different columns of the table. Results within each column are sorted according to AUC.

Marker Combination	Optimal	Linear Terms					Linear + Interaction Terms				
		Large	N = 500		N = 1000		Large	N = 500		N = 1000	
		Sample	Standard	Broad	Standard	Broad	Sample	Standard	Broad	Standard	Broad
1	0.953	0.942	0.771	0.940	0.772	0.941	0.946	0.771	0.944	0.772	0.945
2	0.904	0.894	0.770	0.893	0.771	0.893	0.897	0.770	0.895	0.772	0.894
3	0.899	0.820	0.770	0.817	0.771	0.820	0.822	0.770	0.818	0.771	0.821
4	0.898	0.782	0.769	0.771	0.771	0.773	0.802	0.769	0.770	0.771	0.787
5	0.893	0.775	0.769	0.770	0.771	0.772	0.800	0.769	0.770	0.770	0.772
6	0.893	0.775	0.769	0.769	0.770	0.772	0.786	0.769	0.769	0.770	0.772
7	0.892	0.774	0.769	0.769	0.770	0.770	0.785	0.769	0.769	0.770	0.770
8	0.889	0.774	0.769	0.769	0.769	0.770	0.783	0.768	0.769	0.770	0.770
9	0.888	0.774	0.769	0.769	0.769	0.770	0.780	0.768	0.768	0.769	0.769
10	0.893	0.774	0.768	0.768	0.768	0.768	0.777	0.767	0.748	0.768	0.768

Appendix B: Combinations Gone Wrong

For the bivariate binormal equal correlation scenario, we showed in Section 2.2.1 that the optimal combination of X and Y is $X + \gamma Y$, where $\gamma = \frac{\alpha_2}{\alpha_1} = \frac{\mu_Y - \rho\mu_X}{\mu_X - \rho\mu_Y}$. Using this value of γ always improves performance relative to X alone. However, in practice we generally do not know this optimal γ. One must be cautious when combining markers, as non-optimal values of γ can yield combinations with worse performance than X alone, as shown in Figure B1 below.

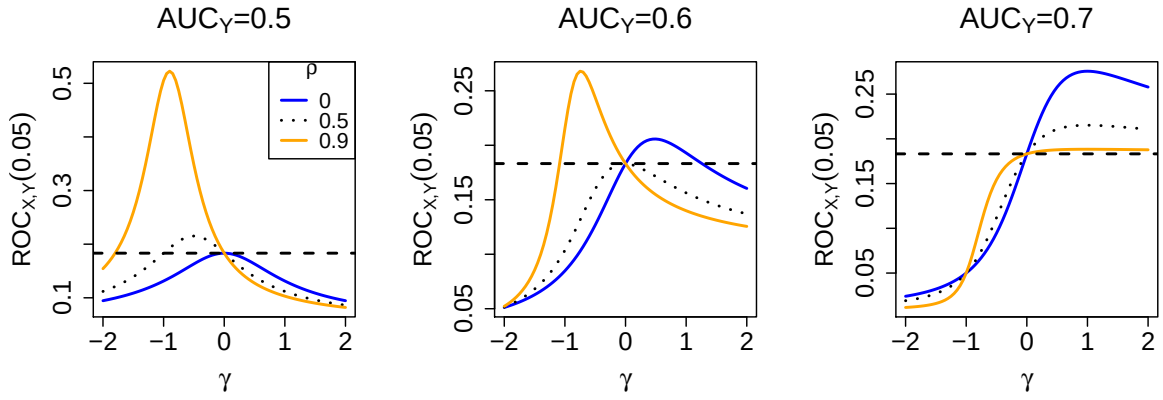
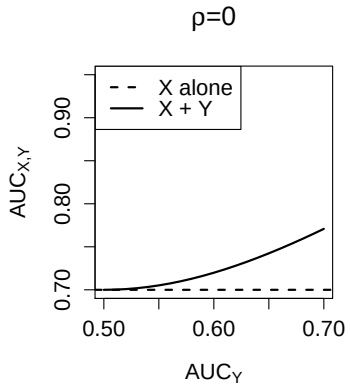
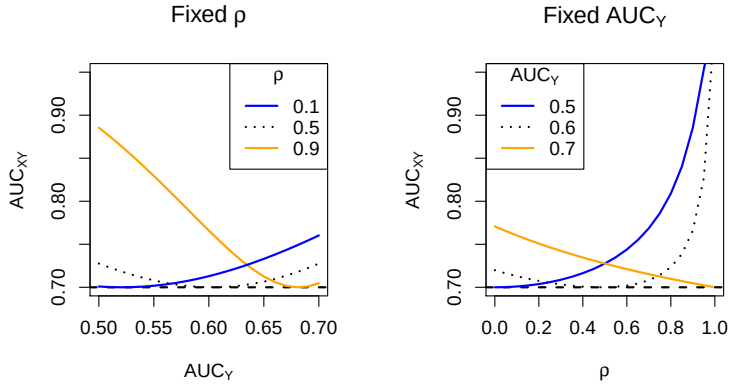


Figure B1: Bivariate Binormal Equal Correlation Markers - Detection of cases by the combination $X + \gamma Y$, using the threshold that leads to FPR = 0.05. The peaks in these plots denote the optimal combination performance, observed when $\gamma = \frac{\alpha_2}{\alpha_1}$. The performance of X alone ($\text{ROC}_X(0.05) = 0.183$) is indicated by a black dashed line.

(a) No Correlation



(b) Equal Correlation



(c) Unequal Correlation

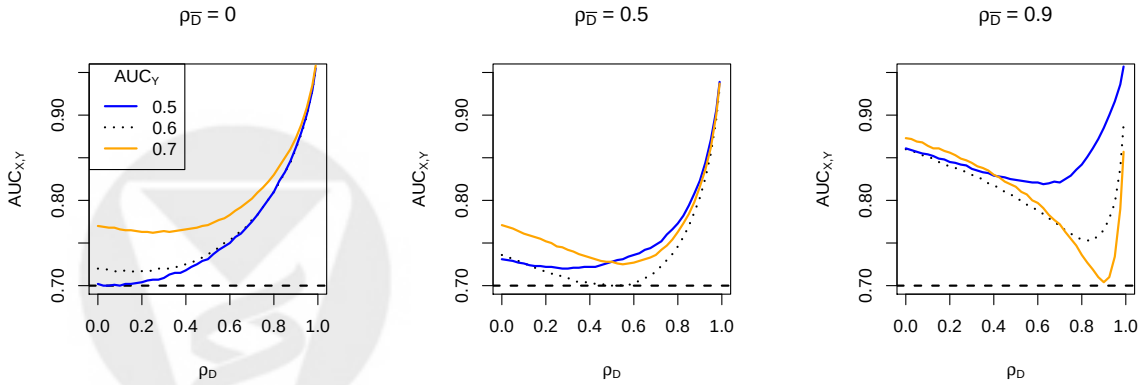


Figure S1: Bivariate Binormal Markers - Using $\text{AUC}_{X,Y}$ as a measure of classification performance of the combination (X,Y). The baseline marker X alone has AUC = 0.7 and is indicated by a black dashed line. Shown are settings where (a) the correlation between markers in both cases and controls is 0, (b) the correlation is equal in cases and controls with positive coefficient ρ, and (c) the markers have a different correlation in controls and in cases, with correlation coefficients $\rho_{\bar{D}}$ and ρ_D, respectively.

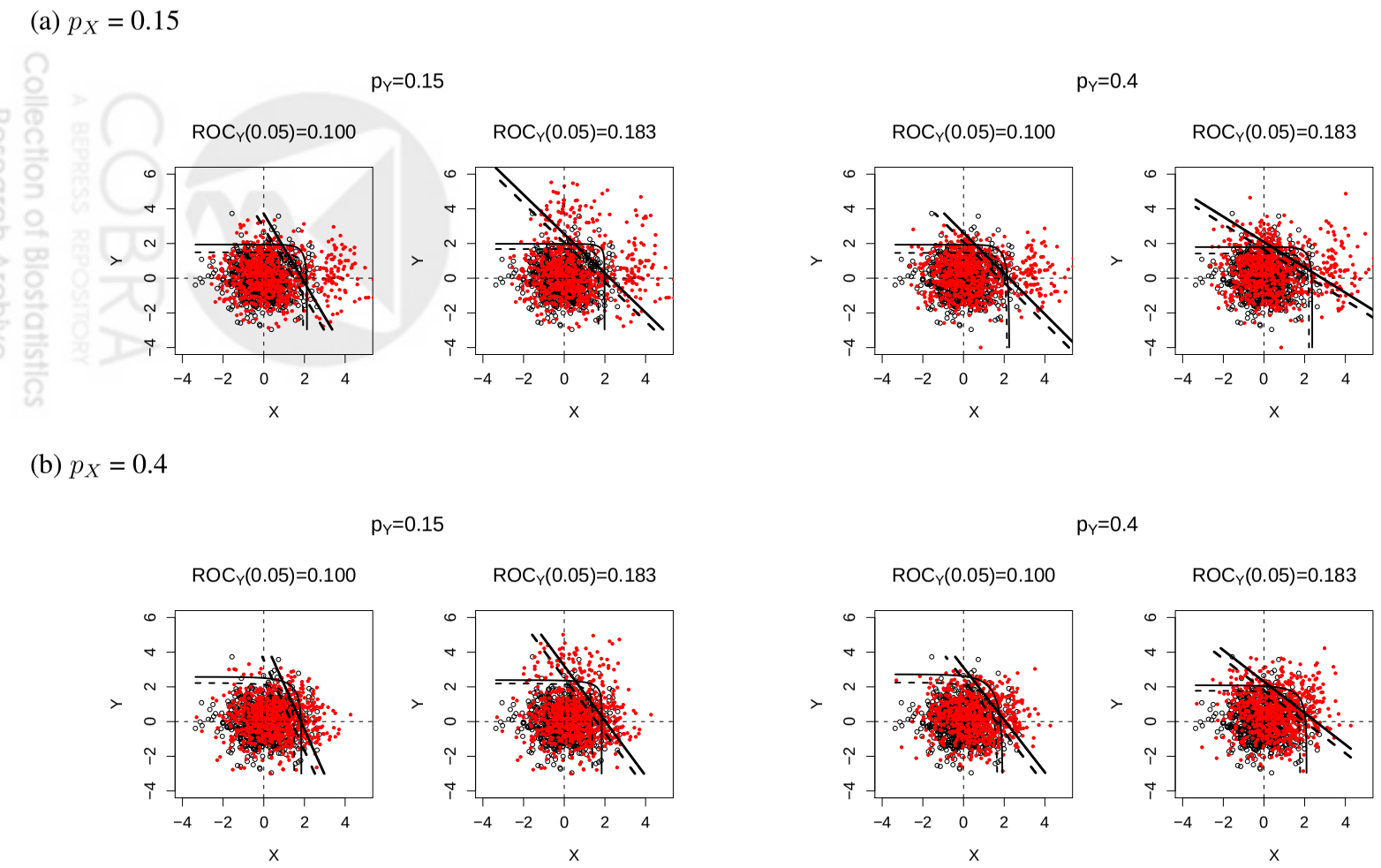
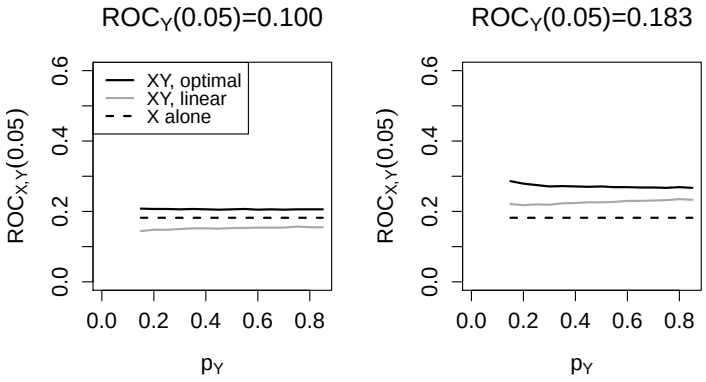


Figure S2: Mixture Binormal Markers - Decision boundaries that separate positive and negative classifications based on (X,Y). X and Y are discriminatory in proportions p_X and p_Y of cases, respectively. We explore different scenarios with p_Y = 0.15, 0.4 and ROC$_Y$(0.05) = 0.100, 0.183 when (a) p_X = 0.15 and (b) p_X = 0.4. The FPR = 0.05 and FPR = 0.10 boundaries are shown with solid and dashed curves, respectively. Both the optimal and linear boundaries are shown. The solid points represent cases, while the hollow circles represent controls.

(a) $p_X = 0.15$



(b) $p_X = 0.4$

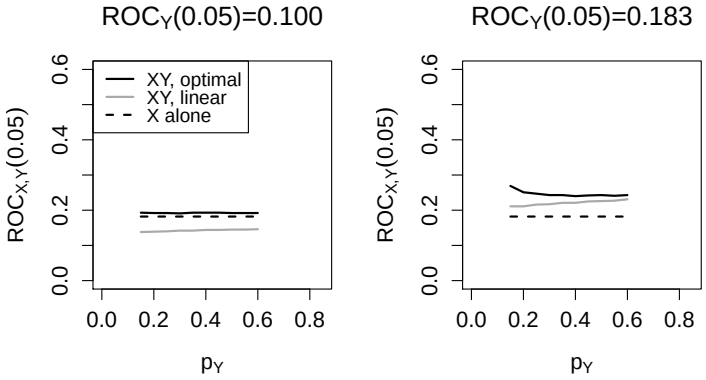


Figure S3: Mixture Binormal Markers - Comparison of decision rules in detecting cases when FPR = 0.05. Shown here are (X,Y) combinations using the optimal risk score and a logistic model with linear terms only, for varying values of p_Y and ROC$_Y$(0.05) when (a) p_X = 0.15 and (b) p_X = 0.4. The baseline marker X alone detects 18% of cases and is indicated by a black dashed line.

Table S1: Results of data analysis for a single dataset derived from the original colon cancer dataset by selecting 100 markers at random and a random split into training (N = 70) and test (N = 70) datasets. Test dataset AUC values for top ranking marker combinations selected using the training dataset are presented here. Marker combinations are different in different columns of the table. Results within each column are sorted according to AUC.

Marker	Linear Terms		Linear + Interaction Terms	
Combination	Standard	Broad	Standard	Broad
1	0.730	0.694	0.791	0.693
2	0.724	0.658	0.767	0.692
3	0.723	0.651	0.728	0.676
4	0.715	0.639	0.725	0.660
5	0.709	0.623	0.719	0.656
6	0.691	0.622	0.708	0.656
7	0.691	0.622	0.690	0.653
8	0.678	0.613	0.683	0.639
9	0.659	0.556	0.666	0.631
10	0.641	0.477	0.642	0.504