BIOMETRICS 58, 657–664
September 2002

CONSULTANT'S FORUM

Combining Several Screening Tests: Optimality of the Risk Score

Martin W. McIntosh[1] and Margaret Sullivan Pepe[1,2,*]

[1]Division of Public Health Sciences, Fred Hutchinson Cancer Research Center,
1100 Fairview Avenue North, Seattle, Washington 98109-1024, U.S.A.
[2]Department of Biostatistics, University of Washington, Seattle, Washington 98195, U.S.A.
*email: mspepe@u.washington.edu

SUMMARY. The development of biomarkers for cancer screening is an active area of research. While several biomarkers exist, none is sufficiently sensitive and specific on its own for population screening. It is likely that successful screening programs will require combinations of multiple markers. We consider how to combine multiple disease markers for optimal performance of a screening program. We show that the risk score, defined as the probability of disease given data on multiple markers, is the optimal function in the sense that the receiver operating characteristic (ROC) curve is maximized at every point. Arguments draw on the Neyman–Pearson lemma. This contrasts with the corresponding optimality result of classic decision theory, which is set in a Bayesian framework and is based on minimizing an expected loss function associated with decision errors. Ours is an optimality result defined from a strictly frequentist point of view and does not rely on the notion of associating costs with misclassifications. The implication for data analysis is that binary regression methods can be used to yield appropriate relative weightings of different biomarkers, at least in large samples. We propose some modifications to standard binary regression methods for application to the disease screening problem. A flexible biologically motivated simulation model for cancer biomarkers is presented and we evaluate our methods by application to it. An application to real data concerning two ovarian cancer biomarkers is also presented. Our results are equally relevant to the more general medical diagnostic testing problem, where results of multiple tests or predictors are combined to yield a composite diagnostic test. Moreover, our methods justify the development of clinical prediction scores based on binary regression.

KEY WORDS: Biomarkers; Diagnosis; Likelihood ratio; Prediction; Prognosis; ROC curve.

1. Introduction

Medical tests for the early detection and diagnosis of disease are a routine part of clinical practice. Recently, research into the development of tumor biomarkers for cancer screening and diagnosis has grown considerably (Henson, Srivastava, and Kramer, 1999; Srivastava and Kramer, 2000), fueled in part by technological advances in genetics and immunology. Tumor biomarkers discovered thus far, such as PSA for prostate cancer and CA 125 for ovarian cancer, while promising, are imperfect. Many diseased subjects have normal tumor marker concentrations, yielding false-negative screening tests, and disease-free subjects often have elevated concentrations leading to unnecessary diagnostic work-up and possible surgery.

One approach to improve the performance of screening with tumor biomarkers uses several markers together. This approach is based on the observation that cancers are heterogeneous with respect to the genes they overexpress, meaning that cancers not expressing one tumor marker may express one or more others (Bast, 1993). This article is concerned with combining results of several tumor biomarker screening tests into a composite screening test. In a more general context, applicable also to diseases other than cancer, we are concerned with the problem of combining multiple medical diagnostic tests to yield a composite diagnostic test that more accurately detects presence of disease.

Combining multiple tests requires an algorithm to classify subjects into one of two groups: those suspected of having the disease and those not suspected. Classification is an old problem in statistics, for which a variety of standard approaches are available (Dudoit, Fridlyand, and Speed, 2000), many of which have already been applied to this problem, including normal linear discriminant analysis (Su and Liu, 1993), regression trees (Woolas et al., 1995), and artificial neural networks (Zhang et al., 1999). Normal linear discriminant analysis identifies the linear combination of test results that best separates the standardized difference in means between diseased and nondiseased subjects. Regression trees, artificial neural net-

work approaches, and other nearest neighbor methods require fewer parametric assumptions than normal discriminant analysis. None of these approaches, however, have been justified as yielding optimal combinations of markers in general for the purposes of medical diagnosis and screening.

In this article, we show that binary regression methods yield optimal scores for combining diagnostic tests. Our arguments draw on signal detection theory developed in the 1950s and 1960s (Green and Swets, 1996; Egan, 1975). In particular, we show that the Neyman–Pearson result (which is familiar to statisticians in the context of hypothesis testing but applies equally well to medical diagnostic testing) implies that the likelihood ratio function of multiple markers is optimal for application to disease screening. We then use Bayes' theorem to show that screening rules based on risk scores (i.e., prediction probabilities of being diseased) are equivalent to those based on likelihood ratio scores and note that the former are easily obtained using standard binary regression methods. The main conclusion, therefore, is that binary regression models predicting disease as a function of diagnostic tests estimate the optimal combination of the tests for classifying a subject as diseased or not. Thus, if discriminant analysis, neural networks, or regression trees perform well, it is only because they, too, are approximating this binary regression. However, binary regression (e.g., logistic regression) has noteworthy advantages relative to other approaches, including well-developed model fitting and inferential procedures, widely available software and procedures for covariate selection that can be useful for marker selection, and the accommodation of case–control sampling in study design.

Sections 2 and 3 give our arguments for the optimality of the risk score and the use of binary regression to estimate it. Section 4 gives practical suggestions for flexible and robust model fitting. This article also contributes, in Section 5, a biologically motivated simulation model for multiple tumor markers. This may be useful for evaluating the screening behavior of any marker selection method, not only that recommended in this article. Section 6 demonstrates large-sample behavior of our recommendation, and Section 7 demonstrates it for combining a newly discovered ovarian cancer marker with CA 125. Some concluding remarks are given in Section 8, including discussion of the classic Bayesian decision framework and how our optimality results contrast with those from that theory.

2. Optimality of the Likelihood Ratio Combination

In the context of disease screening or diagnostic testing, the value of a screening decision rule is measured by its true positive rate, TPR = P[screen positive | diseased], which is also called the sensitivity, and its false-positive rate FPR = P[screen positive | not diseased], which is one minus the specificity. A quality screening test has a high TPR, so that few diseased subjects are missed, and a low FPR, so that most nondiseased subjects are spared unnecessary diagnostic workup and/or treatment. For most tests, a variety of (TPR, FPR) values can be achieved by varying the criterion for defining a positive screening decision. For example, if a single marker, Y, is measured on a continuous scale with higher values more indicative of disease, then higher thresholds, c, used to define a positive result, $(Y > c)$, yield tests with lower TPR and lower FPR. The monotone increasing curve that plots TPR versus FPR for all possible thresholds c is called the receiver operating characteristic (ROC) curve (Hanley, 1989).

To compare two different screening tests, we must do so at comparable FPRs (or equivalently, at comparable TPRs). At a fixed false-positive rate, f_0, the better screening rule is the one achieving the higher TPR. We define the optimal screening rule at f_0 as the one that achieves the maximum possible TPR while maintaining FPR = f_0. It is possible that one screening test is best at some values of FPR but not at others.

Suppose that there are K tests (i.e., markers or other screening tests). We denote the result of the kth test by Y_k, $k = 1, \ldots, K$, and the vector of test results as $\boldsymbol{Y} = (Y_1, \ldots, Y_K)$. Let D be a binary variable denoting disease status, with $D = 1$ for diseased subjects and $D = 0$ for nondiseased subjects. A screening rule uses $\boldsymbol{Y}$ to classify the unobserved D as one or zero. A key result from decision theory, the Neyman–Pearson lemma, states that, for any f_0, the screening rule with the highest TPR based on $\boldsymbol{Y}$ among all possible rules based on $\boldsymbol{Y}$ is the likelihood ratio rule. This gives a positive screening decision whenever

$$LR(\boldsymbol{Y}) > c(f_0), \tag{1}$$

where $LR(\boldsymbol{Y}) = P(\boldsymbol{Y} \mid D = 1)/P(\boldsymbol{Y} \mid D = 0)$ and $c(f_0)$ is chosen so that $f_0 = P\{LR(\boldsymbol{Y}) > c(f_0) \mid D = 0\}$.

This result will be familiar to most statisticians in the context of statistical hypothesis testing (Neyman and Pearson, 1933). Let $D = 0$ and $D = 1$ denote the null and alternative hypotheses, respectively, and $\boldsymbol{Y}$ denote the sample data; the screening test based on $\boldsymbol{Y}$ is the analogue of the rule for rejecting the null hypothesis $(D = 0)$ in favor of the alternative $(D = 1)$. Observe that, in this analogy, type 1 error corresponds to the FPR and statistical power corresponds to the TPR. The Neyman–Pearson result states that the likelihood ratio rule is the uniformly most powerful (UMP) test achieving the highest statistical power among all tests with the same type 1 error rate. Analogously, we can say that the likelihood ratio function of the K screening tests, $LR(\boldsymbol{Y})$, and rules based on its exceeding a threshold achieve the highest TPR possible among all screening tests based on $\boldsymbol{Y}$ with FPR = f_0. We call it the uniformly most sensitive (UMS) screening test based on the combination of $(Y_1, \ldots, Y_K)$.

Optimality of the likelihood ratio rule has long been recognized in signal-detection theory (Green and Swets, 1966; Egan, 1975). Indeed, it was an audiologist who studied under Green (Huanping Dai) that initially brought this theory to our attention. However, the application of the result to combining multiple medical tests has not been emphasized in that literature. Interestingly, Green and Swets and Egan noted that likelihood ratio rules (1) are optimal in two other respects. First, the rule that minimizes the overall misclassification rate is of the form (1) for some f_0. Second, if costs are associated with the two types of errors, false positives and false negatives, then the rule that minimizes expected cost is in the likelihood ratio family of rules. Thus, from several points of view, rules based on the likelihood ratio function exceeding a threshold are simply the best.

In the statistical literature for combining multiple predictors, Baker (1995, 2000) noted optimality of the likelihood ratio score. His arguments draw on cost-effectiveness theory

(Weinstein et al., 1980) rather than on the Neyman–Pearson result. Analogies are drawn between FPR as a measure of cost and TPR as a measure of benefit of the screening program. The optimum benefit given a cost constraint is then provided by the likelihood ratio rule. Baker (2000) uses nonparametric methods to directly approximate the likelihood ratio function. The result in the next section suggests that an alternative approach is to make use of standard binary regression methodology to achieve optimality.

3. Optimality of the Risk Score

Although extremely elegant, the Neyman–Pearson result as presented above appears to require specification of the likelihood ratio function, $LR(\boldsymbol{Y}) = P(\boldsymbol{Y} \mid D = 1)/P(\boldsymbol{Y} \mid D = 0)$, through the determination of its constituent probability distributions, $P(\boldsymbol{Y} \mid D = 1)$ and $P(\boldsymbol{Y} \mid D = 0)$. This can be an onerous task because tumor marker behavior is often irregular and complex multivariate distributions are required to adequately represent them. See, e.g., the model presented in Section 5. Here we show that it is not necessary to specify the constituent distributions because rules based on $LR(\boldsymbol{Y})$ are equivalent to rules based on the risk score $p(\boldsymbol{Y}) = P(D = 1 \mid \boldsymbol{Y})$, which is relatively easy to approximate with binomial regression tools.

Indeed, by Bayes' rule,

$$\begin{aligned} p(\boldsymbol{Y}) &= P(D = 1 \mid \boldsymbol{Y}) \\ &= \frac{P(\boldsymbol{Y} \mid D = 1)P(D = 1)}{\{P(\boldsymbol{Y} \mid D = 1)P(D = 1) + P(\boldsymbol{Y} \mid D = 0)P(D = 0)\}} \\ &= \frac{LR(\boldsymbol{Y})q}{\{LR(\boldsymbol{Y})q + 1\}}, \end{aligned}$$

where $q = P(D = 1)/P(D = 0)$ is the odds of disease in the population. This expression shows that the risk score is a monotone increasing function of $LR(\boldsymbol{Y})$, and so the likelihood ratio rule (1) can be rewritten as

$$p(\boldsymbol{Y}) > c^*(f_0), \tag{2}$$

where the constant $c^*(f_0)$ is chosen to yield FPR $= f_0$, i.e., $c^*(f_0) = c(f_0)q/(c(f_0)q+1)$. In summary, rules of the form (2) are UMS and so achieve the highest possible TPR for every FPR value.

Although at first it appears that cross-sectional cohort sampling must be used to estimate the optimal screening rule, binary regression can also accommodate case–control sampling, which is far more common. Specifically, a logistic form with intercept term $\text{logit}(p(\boldsymbol{Y})) = \beta_0 + h(\beta, \boldsymbol{Y})$ produces rules based on the magnitude of $\text{logit}(p(\boldsymbol{Y}))$, which is also optimal because it is a monotone transformation of the risk score. In case–control sampling, only the intercept, β_0, and not the part involving $\boldsymbol{Y}$, $h(\beta, \boldsymbol{Y})$, is affected by the sampling design (Breslow and Day, 1980). However, the family of screening rules is the same for any value of β_0, so we see that the intercept term is not needed; i.e., rules based on $h(\beta, \boldsymbol{Y})$ are also optimal. Because case–control sampling yields valid estimates of $h(\beta, \boldsymbol{Y})$, case–control designs can be used to derive optimal marker combinations with logistic regression.

Turning now to finding the threshold for the risk score, observe that the threshold is defined by the FPR. In particular, because by definition $f_0 = P\{p(\boldsymbol{Y}) > c^*(f_0) \mid D = 0\}$, we have that $c^*(f_0)$ is the $1 - f_0$ quantile of the risk score $p(\boldsymbol{Y})$ (or of the monotone transformation of it that will be used) in the nondiseased population. Empirical or parametric estimates of the $1 - f_0$ quantile of $p(\boldsymbol{Y})$ can be calculated to approximate $c^*(f_0)$.

4. Practical Considerations for Approximating the Risk Score Rule

Binary regression can estimate the optimal combination, but simply including each marker as a linear term may not suffice. Some consideration should be given to choosing a model form appropriate for markers. We now discuss considerations that are particularly relevant to the disease screening context.

Approximating the optimal screening rules over the entire ROC curve is neither necessary nor practical. We approximate the risk score only for a subregion of the marker space corresponding to false-positive rates in a practically relevant subrange of $(0, 1)$ that we denote by $f_0 \in (L, h)$. Although ideally risk scores and thresholds for very low FPRs would be determined, in finite samples, this is not feasible. The risk score can only be estimated over regions where data are available for cases and controls. Moreover, because thresholds are based on the distribution of $p(\boldsymbol{Y})$ in nondiseased subjects, they can only be estimated within the range of data available for controls. Let L be a reasonable lower limit for the estimable FPRs. For example, using $L = 5/n_{\bar{D}}$, where $n_{\bar{D}}$ is the number of nondiseased observations, acknowledges that, in regions with fewer than five nondiseased subjects, thresholds will not be identified from the data. We choose to restrict risk score estimation to the marker range $A = \{\boldsymbol{Y} : Y_k < y_k(L/K), k = 1, \ldots, K\}$, where $y_k(x)$ is the $(1 - x)$ empirical quantile for Y_k in the controls; i.e., we choose to classify points outside of A as diseased in our screening rules because they are extreme in at least one marker component. This not only makes practical clinical sense but also it can be shown that the false classification of nondiseased subjects due to this convention is no larger than L.

In practice, diagnostic and screening tests have some upper bound on the range of FPRs that can be tolerated in practice. This limit is denoted by h. We choose to also restrict risk score estimation to the marker space region corresponding to adequately low FPRs, below h. One strategy for fitting the risk score model roughly within the data region where the risk score rules have FPRs less than h is to do so in two stages. First, the model is fit over the whole region A. Then, using the fitted values, $\hat{p}(\boldsymbol{Y})$, we select the region where roughly $FPR < h$ as $\{\hat{p}(\boldsymbol{Y}) > \hat{c}^*(h)\}$ and refit the risk score model with the observations that lie in this subregion.

Our final screening rule classifies as diseased all subjects who fall outside of A, as nondiseased all subjects with $\hat{p}(\boldsymbol{Y}) < \hat{c}^*(h)$ and uses the fitted risk score function to classify subjects in the remainder of the marker space using an appropriate threshold that depends on the chosen value for f_0. Observe that one should be conservative about choices of L and h, choosing (L, h) so that it includes all FPRs of interest well within the boundaries. Our procedures are less likely to yield good model fitting at the boundaries. An example of this fitting approach can be found in Sections 6 and 7.

5. A Simulation Model for Tumor Biomarkers

This section proposes a biologically motivated probability model to simulate two ovarian cancer biomarkers. We apply our methodology to simulated data in Section 6. Our model draws on the biological theory that (i) cancers may overexpress none, all, or only a subset of candidate markers; (ii) markers measure the overexpression of tumor specific proteins and so tend to elevate or remain unaffected with the onset of the malignancy; (iii) markers in controls behave homogeneously compared with markers in cases; and (iv) marker concentrations are associated with tumor burden, with more advanced cancers, on average, having higher concentrations.

In the model, associated with each diseased subject are indicators $W_i, i = 1, 2$, with $W_i = 1$ indicating if marker i elevates with the onset of cancer. The fraction of the population having expression pattern $\boldsymbol{W} = (W_1, W_2)$ is $p_{ij} = P(W_1 = i, W_2 = j)$. The odds ratio $\Psi = p_{00}p_{11}/p_{10}p_{01}$ provides a convenient characterization of how much markers complement each other. If $\Psi < 1$, the markers tend to detect different malignancies, whereas if $\Psi > 1$, they tend to detect the same cancers and hence have a degree of redundancy. The marginal probabilities, $p_i = P(W_i = 1)$, $i = 1, 2$, and the odds ratio, Ψ, determine all four frequencies, p_{ij}, of expression patterns.

Cases with expression pattern $\boldsymbol{W} = (i, j)$ have marker concentrations clustered together with mean $\mu_{ij} = E\{\boldsymbol{Y} \mid \boldsymbol{W} = (i, j), D = 1\}$ and variance/covariance structure $\Sigma_{ij} = \text{var}\{\boldsymbol{Y} \mid \boldsymbol{W} = (i, j), D = 1\}$. For convenience, we generate marker values using a bivariate normal distribution conditional on the expression pattern. Because expression patterns are not observed, marker concentrations for cases have the following marginal distribution:

$$P(\boldsymbol{Y} \mid D = 1) = \sum_{i=0,1} \sum_{j=0,1} p_{ij} \times \phi_2(\boldsymbol{Y}; \mu_{ij}, \Sigma_{ij}),$$

where ϕ_2 represents a bivariate normal density. We assume nondiseased subjects have biomarker distributions equal to those of cases with expression pattern $\boldsymbol{W} = (0, 0)$, i.e., normal with mean μ_{00} and variance Σ_{00}.

Figure 1 shows a schematic diagram of our simulation model when configured to mimic the two ovarian cancer markers CA 125 (marker 1) and Her 2/neu (marker 2) when used to distinguish early-stage ovarian cancer from healthy subjects (see Karlan, 1993; Crump et al., 2000). The CA 125 marker is found elevated in the sera of nearly 50% ($p_1 = 0.5$) of all early- stage ovarian cancers, and Her2/neu is found elevated in the sera of nearly 30% ($p_2 = 0.3$) of them. These rates are determined using a specificity near 95% (FPR = 0.05). Without loss of generality, we give a standard normal distribution to each marker in the control group. We assume that markers in cases have twice the standard deviation of controls when they respond. To achieve the required level of sensitivity at FPR = 0.05, we assume markers 1 and 2 elevate to values that are, on average, 1.68 and 2.54, respectively.

We separately choose the association parameters as follows. We assume markers are complementary, ($\Psi = 1/10$), which means fewer cases will fall in the upper right ellipse (both markers respond), and so one marker can find many cancers not found by the other. When both markers respond, we give them a mild correlation ($\rho = 0.6$) that is induced by the markers measuring the same tumor burden, but for all other subpopulations of cases, markers are assumed independent. Such independence is justified for ovarian cancer markers based on the findings of Crump et al. (2000), who found that each of five ovarian cancer markers behave independently.

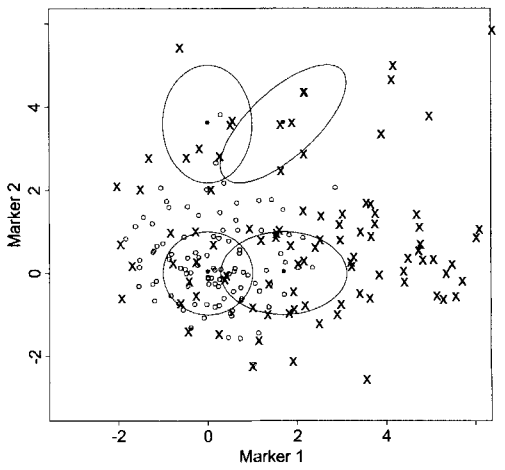

Figure 1. Schematic representation of simulation model along with sample data for 100 cases (denoted by 'o') and 100 controls (denoted by '×'). The lower left ellipse represents a 68% probability region (1 SD) for controls whereas cases are represented by the four ellipses depending on their expression pattern with none, one, or both markers overexpressed. The proportion of cases with various expression patterns are $p_{00} = 0.25$ (no overexpression), $p_{10} = 0.45$ (marker 1 only), $p_{01} = 0.25$ (marker 2 only), $p_{11} = 0.05$ (both markers). The marginal results are chosen to reflect the behavior of two known ovarian cancer markers.

By separating the model's marginal behavior parameters from the association parameters (Ψ, ρ), we can conveniently vary only the association parameters and examine the behavior of combining markers having different capacities when used together but that behave the same when used alone.

6. Large-Sample Results

With our tumor marker model, the exact likelihood ratio function is

$$LR(\boldsymbol{Y}) = \frac{\sum_{i=0,1} \sum_{j=0,1} p_{ij} \times \phi_2\left(\boldsymbol{Y}; \mu_{ij}, \Sigma_{ij}\right)}{\phi_2\left(\boldsymbol{Y}; \mu_{00}, \Sigma_{00}\right)}.$$

The contours shown in Figure 2a represent levels of equal likelihood ratio or, equivalently, of equal cancer risk. Darker levels imply increasing density of cases compared with controls. Observe that the cancer risk is monotone in any direction of increased marker concentration. These contours represent UMS rules and show the optimal boundaries for defining positive screening tests based on (Y_1, Y_2).

The ROC curves for the two markers (alone and combined) are shown in Figure 2b over a range of FPRs in $(0.00, 0.15)$. These curves were generated from one million marker pairs and so represent large-sample results, with statistical uncertainty eliminated. The highest ROC curve represents the optimal screening rule generated from the contours of Figure 2a. The two lowest ROC curves represent the lone use of marker

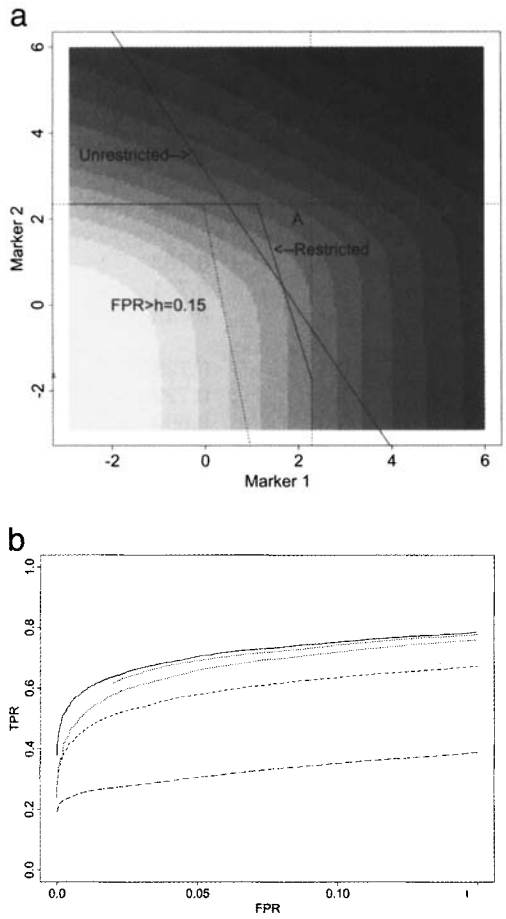

Figure 2. Contours of equal likelihood ratio or equivalently cancer risk (**a**) and receiver operating characteristic curves (ROC) for each marker alone and for combinations (**b**). The ROC curves represent, from lowest to highest, marker 2 alone, marker 1 alone, the unrestricted logistic regression score, the restricted linear regression score, and the optimal risk score.

2 and marker 1, from bottom up, respectively. As the optimal ROC curve indicates, performance can potentially be improved substantially by using both markers together.

Consider now approximating the risk score. A rule based on a maximum likelihood approximation to $\text{logit}(p(\boldsymbol{Y}))$ = $b_0 + b_1 Y_1 + b_2 Y_2$ on the unrestricted marker space is labeled in Figure 2a. Lines parallel to the one shown, $c = b_1 Y_1 + b_2 Y_2$, define screening rules having different false-positive rates, depending on c, where points above the line are classified as screen positive. That ROC curve generated by varying c is third highest shown in Figure 2b. Although this unrestricted linear fit improves performance over those for the single markers, its operating characteristics are substantially less than optimal.

We next restrict attention to the subregion A defined using $L = 0.02$. The border of region A is formed by the vertical and horizontal line segments in Figure 2a. Regions outside the lower left quadrant formed by these two lines are always associated with a screen-positive result. We next fit the linear model using only the observations in region A, then eliminate all subjects having a predicted cancer risk below the top $h = 15\%$ of all nondiseased subjects. These correspond to all observations falling below the dotted-line segment in Figure 2a. The screening rule in this region of the marker space is screen-negative. We then refit the regression once again to the remaining observations to yield the model $p(Y) = b_0' + b_1' Y_1 + b_2' Y_2$. The resultant screening rules are such that points falling outside region A screen positive, those falling below the dotted line segment screen negative, and all remaining points screen positive if they are above $c = b_1' Y_1 + b_2' Y$. One example of the screening rule is shown by the solid piecewise-linear region in Figure a. Varying c produces the ROC curve over the restricted region of FPRs, $(L, h) = (0.02, 0.15)$, which is the second highest ROC curve in Figure 2b. We see that it is substantially closer to the optimal ROC curve than is that derived from the unrestricted linear fit.

Figure 2 indicates that, by fitting the linear logistic model in the restricted space, the estimated risk score well approximates the true risk score in that region. We found this conclusion to hold for a wide variety of simulation model configurations. We used the simulation model described earlier but varied the association parameters and kept the marginal parameters fixed as before. Table 1 shows ROC curve summary statistics for the various configurations examined. The summary statistic displayed is the partial area under the ROC curve (pAUC) between false-positive rates of 0.02 and 0.15 normalized by the maximum value that the partial area can attain for a perfect marker that completely discriminates cases from controls, namely $0.15 - 0.02 = 0.13$. Thus, a normalized pAUC value of 0.70 implies that 70% of the maximum achievable performance was achieved.

Across all configurations, all methods that combine markers perform better than using any single marker on its own. In general, the rule that is linear over the restricted space appears to perform almost as well as the optimal rule and substantially better than the linear score fit on the unrestricted space. We found that combining markers finds its greatest improvement over a single marker when markers complement each other (i.e., when $\Psi < 1$) and when the correlation is low (i.e., when ρ is small). On the other hand, when markers are highly redundant (i.e., when $\Psi = 10$), combining markers by even the optimal method does not find a dramatic improvement over using only the single best marker.

7. Small-Sample Example

Figure 3a shows marker pair concentrations from 51 ovarian cancer cases and 50 controls with benign ovarian disease. One marker represents CA 125 and the other represents a novel marker candidate (McIntosh et al., unpublished manuscript). Both markers are log transformed and then rescaled and recentered so that each marker has zero mean and unit variance in the control group. Imposed over the scatterplot are screening rules with FPR = 0.10 from both an unrestricted fit (long dashed line) and a restricted fit (solid line segments).

Because of the small number of control subjects in this study, we evaluate these markers over a range of FPR between $L = 0.04$ and $h = 0.50$. We see from the ROC curves that CA 125 performs better than the candidate marker when used alone, and combining them together with an unrestricted regression does little better than using only CA 125. However, a restricted regression finds that their combination performs

Table 1
Normalized partial area under the curve for five different families of screening rules calculated for different parameterizations of the simulation model

Ψ	ρ	Marker 1	Marker 2	Unrestricted[a]	Restricted[b]	Optimal[c]
1/10	0	0.649	0.341	0.740	0.779	0.784
1/10	0.5	0.649	0.335	0.735	0.769	0.772
1/10	0.9	0.636	0.344	0.721	0.753	0.756
1/5	0	0.674	0.336	0.727	0.758	0.762
1/5	0.5	0.639	0.348	0.719	0.748	0.754
1/5	0.9	0.636	0.344	0.709	0.738	0.743
1	0	0.644	0.345	0.700	0.724	0.729
1	0.5	0.649	0.342	0.696	0.716	0.722
1	0.9	0.642	0.332	0.679	0.693	0.699
5	0	0.637	0.349	0.691	0.706	0.709
5	0.5	0.647	0.335	0.664	0.678	0.684
5	0.9	0.640	0.348	0.664	0.662	0.675
10	0	0.639	0.335	0.681	0.694	0.700
10	0.5	0.639	0.342	0.667	0.677	0.685
10	0.9	0.643	0.343	0.665	0.669	0.673

[a] Linear fit over unrestricted marker space.
[b] Linear over restricted region.
[c] True risk score contours.

better than each individually. The top ROC curve in Figure 3b indicates that the restricted region linear score provides somewhat higher discriminant capability.

8. Discussion

The key observation in this article is that, for classifying disease status on the basis of multiple predictors (biomarkers or diagnostic tests), the optimal rules are based on thresholds for the risk score. The arguments we have presented here essentially appeared 40 years ago in the literature on signal-detection theory. Green and Swets (1966) noted not only the optimality of the likelihood ratio function but also the relationship between the risk score and the likelihood score. However, these results and their practical consequences have not been well appreciated by statisticians heretofore.

We have emphasized binary regression methods to estimate risk scores. Other approaches are possible. In particular, Bayesian methods have become popular in cancer screening. Sophisticated modeling of cancer incidence along with models for biomarker trajectories after disease onset can be employed to derive estimates of risk given data on multiple markers. To the extent that such methods yield valid estimates of the risk score function (or of monotone functions of it), our arguments indicate that rules based on them are optimal. Alternatively, nonparametric or semiparametric binary regression methods, such as regression trees, could be used to estimate the risk scores. Baker's approach (Baker, 2000), which estimates the likelihood ratio function, can be viewed as estimating a monotone function of the risk score in a nonparametric fashion. It would be interesting to determine if his ordered algorithms induce nonparametric monotone estimates of the risk score function in multiple dimensions.

With Bayesian principles, the optimality of the risk score has been derived using the criterion of minimizing an expected loss function. Given a prior probability for disease, $P(D = 1)$, and a loss function (or cost function) associated with errors in the decision rule, it is well known that decision rules based on the posterior probability of $P(D = 1|Y)$ minimizes the expected loss. Thus, in order to minimize an expected loss, Bayesians already know that the risk score is optimal. See, e.g., Ripley (1996) or McLachlan (1992).

Our approach, in contrast, achieves a different but powerful optimality criterion, one that does not require soliciting prior probabilities or a loss function but simply an acceptable false-positive rate. The risk score maximizes the true positive rate simultaneously for each false-positive rate. The implications of this result are very important in practice. Heretofore, there was nothing in the frequentist framework other than convenience and intuition to argue for using the risk score for combining tumor markers or other diagnostic tests for classification. The intuitive appeal of the risk score is now justified with theory.

We have suggested some approaches to approximating the risk score that reduce the scope and complexity of the task. In particular, we note that the risk score need only be estimated over a subregion of the data space that includes a portion of controls (at the upper but not extreme end of the distribution) and cases that fall in that same region. Our analyses (real and simulated) indicate that simple forms for the risk score are often sufficient in the restricted region and hence that simple rules for screen positivity are optimal in the range of FPRs that is of practical interest.

Binary regression methods are routinely used in practice to develop optimal diagnostic or prognostic scores. This article provides a justification for their use with adequate attention to model fitting since we show that (in large samples) such

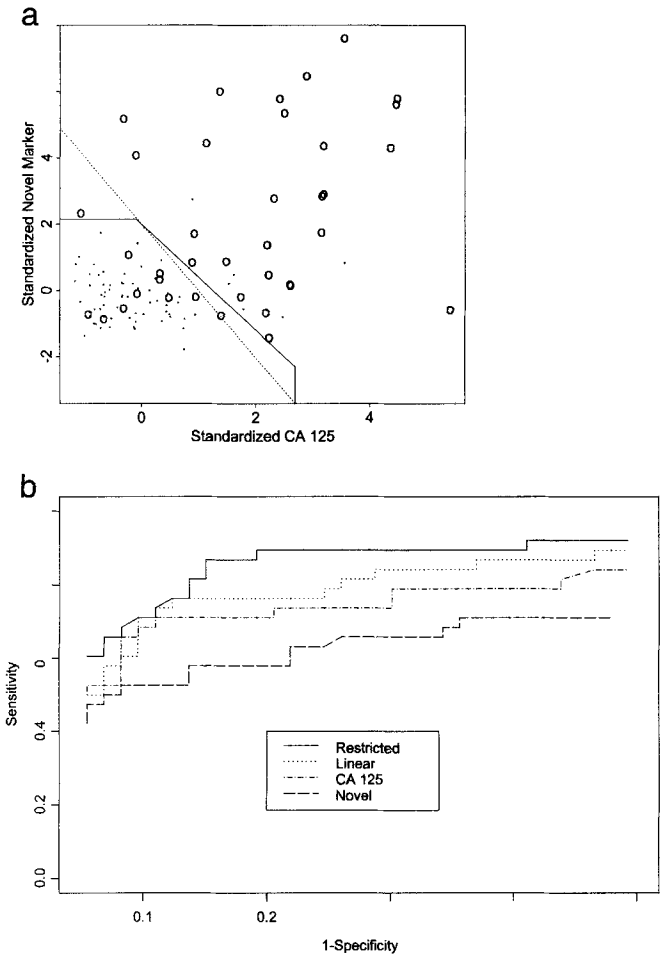

Figure 3. CA 125 and a new candidate marker in 51 cancer cases and 50 healthy controls along with simple linear and piecewise ($L = 0.04, h = 0.50$) screening rules (**a**) and ROC curves (**b**). Shown in (**a**) are the rules corresponding to an estimated FPR = 0.10. In (**b**) are empirical ROC curves for single markers and combinations. Combinations are (i) linear score fit over unrestricted region and (ii) borders and linear score fit over restricted region.

scores are optimal. Small sample behavior relative to other procedures has not been addressed by us here. When combined with the notion of restricting the region of the predictor space for model fitting, sampling variability and its implications for overfitting models to data become major concerns. Some further research into this and into variable selection procedures would be of interest. Model fitting procedures based on analysis of the ROC curves might be considered as an alternative to procedures based on maximum likelihood, for example. Another question relates to the impact of using matching variables in the design of case–control studies of biomarkers, particularly if this results in a distortion of the distribution of matching variables relative to diseased and nondiseased populations.

Acknowledgements

Support for this research was derived from research grants GM-54438 and CA-83636. Our thanks to Steve Neely and Huanping Dai of the BTNRH for stimulating discussions and to Lian Schmidt and Noelle Noble, who helped prepare the manuscript.

Résumé

Le développement des marqueurs biologiques dans le dépistage du cancer est un domaine de recherche très actif. Alors que plusieurs marqueurs existent, aucun isolément n'est suffisamment sensible ou spécifique pour le dépistage d'une population. Nous étudions comment combiner plusieurs marqueurs pour optimiser les performances d'un programme de dépistage. Nous démontrons que le score de risque, défini comme la probabilité d'être atteint sachant les valeurs de différents marqueurs, est une fonction optimale dans le sens où la courbe de ROC est maximisée à chaque point (Argumentation présente dans le lemme de Neyman-Pearson). Ceci contraste avec les critères d'optimalité classiques dans la théorie de la décision qui sont utilisés dans un cadre bayésien et sont fondés sur la minimisation d'une fonction de coût associée aux erreurs de décision. Notre critère d'optimalité est définit dans un cadre purement fréquentiste et n'est pas lié au coût des erreurs de classement. Pour l'analyse les méthodes de régression binaire peuvent être utilisées pour fabriquer les poids relatifs des différents marqueurs au moins sur des grandes séries. Nous proposons quelques modifications des méthodes standard de régression binaire pour qu'elles soient applicables au problème de dépistage d'une maladie. Une évaluation de notre méthode est présentée sur une étude de simulations à partir d'un modèle biologique de marqueurs dans le cancer. Une application à des données réelles de marqueurs du cancer de l'ovaire est également présentée. Nos résultats sont généralisables à tous les problèmes de test diagnostic, où les résultats de plusieurs tests ou facteurs prédictifs sont combinés pour former un test diagnostic composite. De plus notre méthode justifie le développement de scores cliniques fondés sur la régression binaire.

References

Baker, S. (1995). Evaluating multiple diagnostic tests with partial verification. *Biometrics* **51,** 300–337.

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

Bast, C. J. (1993). Perspectives on the future of cancer markers. *Clinical Chemistry* **39,** 2444–2451.

Breslow, N. and Day, N. (1980). *Statistical Methods in Cancer Research, The Analysis of Case–Control Studies,* Volume I. Publication No. 32. Lyon: IARC Scientific.

Crump, K. C., McIntosh, M. W., Urban, N., Anderson, G., and Karlan, B. Y. (2000). Ovarian cancer tumor marker behavior in asymptomatic healthy women: Implications for screening. *Cancer Epidemiology, Biomarkers and Prevention* **9,** 1107–1111.

Dudoit, S., Fridlyand, J., and Speed, T. (2000). *Comparison of discrimination methods for the classification of tumors using gene expression data.* Technical Report 576, Department of Statistics, University of California, Berkeley.

Egan, J. P. (1975). *Signal Detection Theory and ROC Analysis.* New York: Academic Press.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics.* New York: Wiley: New York.

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

Henson, D. E., Srivastava, S., and Kramer, B. S. (1999). Molecular and genetic targets in early detection. *Current Opinion in Oncology* **11,** 419–425.

Karlan, B. Y. (1993). Screening for ovarian cancer: What are the optimal surrogate endpoints for clinical trials? *Journal of Cell Biochemistry* **23,** 227–232.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition.* New York: Wiley.

Neyman, J. and Pearson, E. S. (1933). On the problem of the most efficient tests of statistical hypothesis. *Philosophical Transactions of the Royal Society of London, Series A* **231,** 289–337.

Ripley, B. D. (1996). *Pattern Recognition and Neural Networks.* Cambridge: Cambridge University Press.

Srivastava, S. and Kramer, B. S. (2000). Early detection cancer research network. *Laboratory Investigation* **80,** 1147–1148.

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

Weinstein, M. C., Fineberg, H. V., Elstein, A. S., et al. (1980). *Clinical Decision Analysis.* Philadelphia: W. B. Saunders.

Woolas, R. P., Conaway, M. R., Xu, F., et al. (1995). Combinations of multiple serum markers are superior to individual assays for discriminating malignant from benign pelvic masses. *Gynecologic Oncology* **59,** 111–116.

Zhang, Z., Barnhill, S. D., Zhang, H., Xu, F., Yu, Y., Jacobs, I., Woolas, R. P., Berchuck, A., Madyastha, K. R., and Bast, R. C., Jr. (1999). Combination of multiple serum markers using an artificial neural network to improve specificity in discriminating malignant from benign pelvic masses. *Gynecologic Oncology* **73,** 56–61.

Received April 2001. *Revised February* 2002.
Accepted March 2002.