*Biostatistics* (2003), **4**, 4, *pp.* 523–538
*Printed in Great Britain*

# Combining biomarkers to detect disease with application to prostate cancer

RUTH ETZIONI*, CHARLES KOOPERBERG, MARGARET PEPE, ROBERT SMITH

*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA, USA*
retzioni@fhcrc.org

PETER H. GANN

*Department of Preventive Medicine, Robert H. Lurie Comprehensive Cancer Center, Northwestern University Medical School, Chicago, IL, USA*

## SUMMARY

In early detection of disease, combinations of biomarkers promise improved discrimination over diagnostic tests based on single markers. An example of this is in prostate cancer screening, where additional markers have been sought to improve the specificity of the conventional Prostate-Specific Antigen (PSA) test. A marker of particular interest is the percent free PSA. Studies evaluating the benefits of percent free PSA reflect the need for a methodological approach that is statistically valid and useful in the clinical setting. This article presents methods that address this need. We focus on and-or combinations of biomarker results that we call logic rules and present novel definitions for the ROC curve and the area under the curve (AUC) that are applicable to this class of combination tests. Our estimates of the ROC and AUC are amenable to statistical inference including comparisons of tests and regression analysis. The methods are applied to data on free and total PSA levels among prostate cancer cases and matched controls enrolled in the Physicians' Health Study.

*Keywords*: AUC; Classification; Logic regression; ROC.

## 1. INTRODUCTION

In early detection of disease, combinations of biomarkers promise improved diagnostic performance over single markers, which may be lacking in sensitivity and/or specificity. An important example is that of prostate cancer screening with Prostate-Specific Antigen (PSA). Although high PSA levels are associated with prostate cancer, benign conditions may also cause elevation of PSA. Consequently, the conventional criterion for a positive test (PSA $> 4.0$ ng ml$^{-1}$) yields a non-trivial false-positive rate, with high costs in terms of unnecessary biopsies and emotional distress (Brawer, 2000). However, PSA consists of two different subtypes, free and complexed PSA, and while their sum tends to rise in the presence of a malignancy, the proportion of free PSA tends to decline (Stenman *et al.*, 1991). This fact leads to the obvious question: Could combining information on the ratio of free to total PSA (RPSA) with the total PSA level (TPSA) improve discrimination of prostate cancer cases from healthy men?

*To whom correspondence should be addressed.

There is a substantial literature on the potential gains that might be achieved from the use of RPSA in combination with TPSA (e.g. Beduschi and Oesterling, 1998; Brawer, 2000; Carlson *et al.*, 1998; Catalona *et al.*, 1995, 1997; Gann *et al.*, 2002; Partin *et al.*, 1996; Reissigl *et al.*, 1996). To date, the vast majority of published studies have explored the use of RPSA when TPSA levels are mildly elevated, within a specified diagnostic gray zone, or reflex range. The combination tests are of the form: test positive if $R(a, b, c)$ is true, where

$$R(a, b, c) = \text{TPSA} > c \quad \text{OR} \quad (b < \text{TPSA} \leqslant c \quad \text{AND RPSA} < a).$$

Here the reflex range is the interval $(b, c)$.

The endpoints of the TPSA reflex range and the optimal threshold for RPSA within this range are matters of some debate. Initial studies focused on a reflex range for TPSA of 4–10 ng ml$^{-1}$, and found that use of RPSA within this range appeared to substantially improve specificity with only small losses in sensitivity (Catalona *et al.*, 1995; Partin *et al.*, 1996). Subsequent studies suggested that RPSA might be useful when PSA levels were even lower than 4.0 ng ml$^{-1}$ (Catalona *et al.*, 1997; Reissigl *et al.*, 1996). A recent report by Gann and colleagues observed that use of RPSA within a TPSA reflex range of 3–10 ng ml$^{-1}$ could actually improve both specificity and sensitivity simultaneously relative to the conventional test (Gann *et al.*, 2002). As with the reflex range, the recommended percent free PSA threshold has varied across studies, from a minimum of approximately 10% to a maximum of 25%.

Although the studies published to date collectively suggest that combining RPSA with TPSA may indeed be useful, the individual studies each explore a very restricted subspace of the potential combination rules (e.g. by fixing the reflex range) and generally do not quantify the statistical significance of any apparent improvement in diagnostic performance. In our opinion this is largely a consequence of the lack of an accessible statistical methodology for identifying and comparing tests that combine information on multiple markers in a clinically meaningful way. In this article we introduce such a methodology and illustrate its utility in practice. We consider the space of clinically meaningful combination rules to be the set of 'and–or' combinations of threshold rules in each biomarker, which we refer to as *logic rules*; the rule $R(a, b, c)$ above is a specific instance of a logic rule.

Logic combination rules are preferred by clinicians for their interpretability and simplicity. By identifying the space of candidate rules with the logic rules, we consider combination tests that have not been previously explored. These include, for instance, rules that extend the TPSA reflex range below 4.0 ng ml$^{-1}$, but use a more stringent RPSA criterion for men below versus above this threshold; such a rule has been suggested in the clinical literature, but never formally evaluated (Beduschi and Oesterling, 1998). We also introduce graphical and quantitative methods for comparing the performance of the combination rules with conventional or competing rules. These methods extend Receiver Operating Characteristic (ROC) curve methodology in an intuitive and interpretable way from tests based on a single marker to tests based on logic combinations of markers. The methods are applied to data on free and total PSA levels among prostate cancer cases and matched controls participating in the Physicians' Health Study (Gann *et al.*, 1995, 2002). By introducing these methods and illustrating their applicability to a highly relevant and controversial problem, we hope to provide a useful approach to developing combination tests for the early detection of disease.

## 2. METHODS

### 2.1 *Overview*

In this section, we develop a definition of the ROC curve for logic combinations of biomarkers ('logic rule ROC curve'). This development assumes availability of a classification algorithm for identifying predictive rules combining multiple biomarkers. Classification algorithms search the space of possible

rules for the optimal rule, namely the one minimizing a specified objective function. After presenting our notation, we describe the classification algorithm used in our application. The definition of the logic rule ROC curve follows from assuming a particular form for the objective function. This objective function, which is closely related to the probability of misclassification, is applicable also in the case of a single biomarker, and provides us with a unified definition of the ROC curve which is valid for both the single- and multiple-biomarker settings. For comparing rules we define a concept analogous to the area under the ROC curve (AUC) with one marker, which we call the Probability of Correct Classification (PCC). The PCC is relatively straightforward to estimate in practice and lends itself to statistical inference, in contrast to the empirical AUC. We derive conditions under which the PCC equals the empirical AUC, which allows us to determine how closely the PCC will approximate the area under the logic rule ROC curve in practice. We then show how to apply theory recently developed for AUC regression with one marker (Dodd and Pepe, 2002a) to the PCC to compare the performance of different rules while adjusting for covariates.

### 2.2 *Notation*

Consider the case of two markers, $X$ and $Y$; the methods presented below extend easily to the case of more than two markers. In our application, $X$ will denote TPSA and $Y$ RPSA. Denote the marker values for cases by $X^D$ and $Y^D$ and for controls by $X^{\bar{D}}$ and $Y^{\bar{D}}$. Let $c_1, c_2, \ldots c_m$ be the set of thresholds of interest in $X$, and, similarly, let $d_1, d_2, \ldots d_n$ be the set of thresholds of interest in $Y$; the sets $\{c_i\}$ and $\{d_j\}$ may be informed by the specific application or they may simply be evenly spaced percentiles or uniform grids over the relevant biomarker ranges.

The set of logic rules consists of the space of rules given by and–or combinations of expressions like $X > c_i$ and $Y > d_j$. Graphically, a logic rule would be represented by a set of rectangles, or step-functions traversing the scatterplot of $X$–$Y$ values as illustrated in Figure 1. Many classification algorithms exist to identify 'good' logic rules (e.g. Breiman *et al.*, 1984; Quinlan, 1993; Ruczinski *et al.*, 2003). These algorithms typically require specification of an objective function which quantifies the predictive performance of any given rule. To identify high-quality rules, the algorithms search through the space of classification rules to find the rules that optimize the objective function. We use a specific classification algorithm, and a specific objective function, but the development here does not depend on these particular choices.

### 2.3 *Logic regression*

The classification algorithm used in this application is logic regression, an adaptive regression methodology developed for binary covariates (Ruczinski *et al.*, 2003). Logic regression searches for Boolean combinations of predictors in the entire space of such combinations, while being completely embedded in a regression framework, where the quality of the models is determined by the respective objective functions of the regression class.

As in many nonparametric regression methodologies, the goal in logic regression is to predict a response variable based on predictor variables. We here assume that all predictors are binary. In our setting, the predictors are logic combinations of threshold conditions in each biomarker (e.g. $X > c_i$ or its complement). The type of regression problem is determined by an objective function that relates fitted values with the response. Possible objective functions include the residual sum of squares in linear regression, the log-likelihood in generalized regression, the partial log-likelihood in Cox regression, or misclassification in classification problems. In our application we use classification models; the response is an indicator of case-control status and our objective function is simply the number of misclassified individuals.

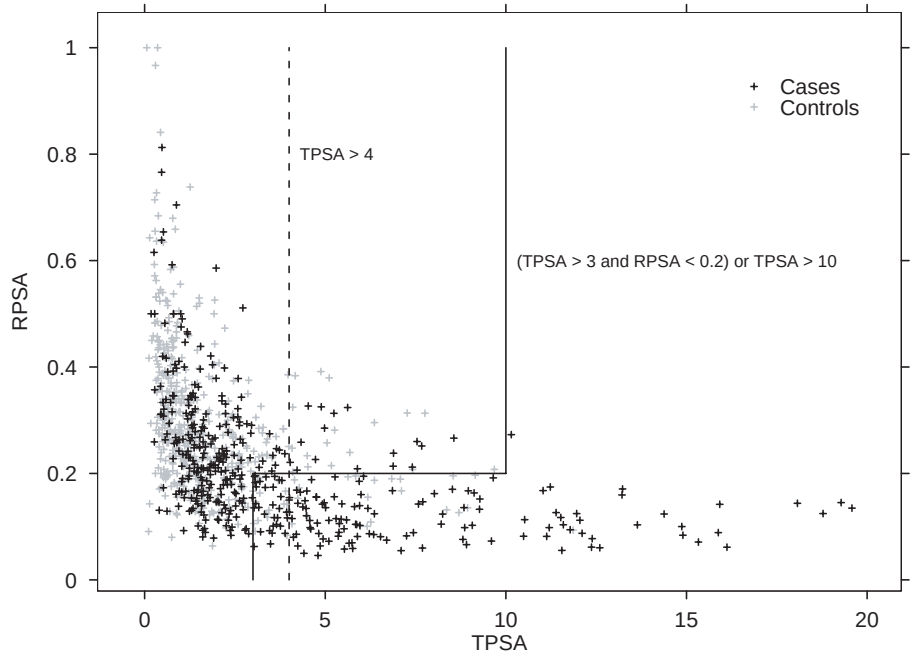


Fig. 1. Graph of the TPSA/RPSA data for study participants, together with the conventional TPSA-based rule and the rule identified by Gann *et al.* (2002). For display purposes we have plotted data from a random 50% of controls, and have cut off the horizontal axis at TPSA $= 20$ ng ml$^{-1}$; 17 cases and 8 controls had TPSA values above 20 ng ml$^{-1}$. All of these cases and 6 of the 8 controls also had RPSA values below 0.2.

In logic regression, the challenge is to find good candidates for the logic term, as the collection of all Boolean expressions is enormous. Using a tree-like representation for logic expressions, we can adaptively select this term using a simulated annealing algorithm. In our setting leaves of each tree are the threshold conditions in each biomarker, and the root and knots of the tree are the Boolean (and–or) operators (Figure 2). Simulated annealing is a probabilistic, iterative algorithm. At each step a possible operation on the current tree, such as adding or removing a knot, is proposed at random. This operation is always accepted if the new logic tree has a better score (objective function value) than the old logic tree, otherwise it is accepted with a probability that depends on the difference between the scores of the old and the new tree and the stage of the algorithm. Properties of the simulated annealing algorithm depend heavily on Markov chain theory and thus on the set of operations that can be applied to logic trees (van Laarhoven and Aarts, 1987). The complexity of a specific model is defined by the size of its logic tree which is given by the number of leaves. Naturally, models of greater complexity will tend to fit the observed data best. To avoid overfitting, the algorithm first selects model size using a cross-validation approach (Ruczinski *et al.*, 2003).

### 2.4 *The logic rule ROC curve*

The misclassification objective function can be written as $FP + FN$, where $FP$ denotes the number of false-positive errors and $FN$ the number of false-negative errors. More generally, a weighted misclassification error function, $L(\alpha) = \alpha FP + FN$, may be more appropriate, for example because the ratio of cases to controls in the data set may be arbitrary. Higher values of $\alpha$ will lead the algorithm to search for logic

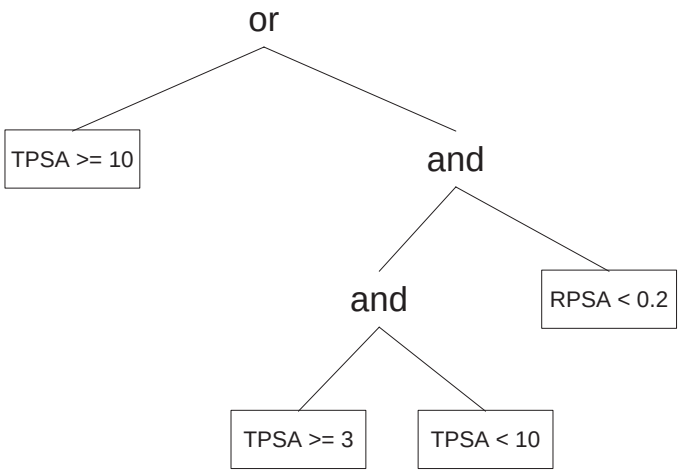


Fig. 2. An example of a logic tree, corresponding to the rule: TPSA $\geqslant$ 10.0 ng ml$^{-1}$ OR ((3 $\leqslant$ TPSA $<$ 10) AND RPSA $<$ 0.2). (Note that this rule would usually be reduced to TPSA $\geqslant$ 10.0 ng ml$^{-1}$ OR (3 $\leqslant$ TPSA AND RPSA $<$ 0.2) by the logic regression algorithm.)

rules with lower false-positive error rates and conversely. The weight $\alpha$ may in general be thought of as an index of conservatism with higher values of $\alpha$ yielding rules that are more conservative in the sense that they are less likely to declare a test positive. The logic rule ROC curve is defined in terms of $\alpha$ as follows.

DEFINITION 1 The logic rule ROC curve, $ROC(\alpha)$, is a plot of the true- versus false-positive rates corresponding to the optimal logic rules under the weighted misclassification error function, $L(\alpha)$, as $\alpha$ varies.

This definition is justified also in the one-dimensional (single marker) setting. In the single-marker case, the ROC curve, $ROC(c)$, corresponding to the test TPSA $>$ c, consists of the pairs of true- and false-positive rates as $c$ varies. However, suppose that false-positive errors have weight $\alpha$ relative to false-negative errors. Then, for any given value of $\alpha$, there will exist a threshold $c$ that minimizes the weighted misclassification error function, $L(\alpha)$. As $\alpha$ increases, so will the cutoff $c$ and vice versa. Since $c$ is therefore a monotonic function of $\alpha$, we can define the ROC curve as the plot of true-positive versus false-positive rates either as the cutoff $c$ changes or as the index $\alpha$ changes. This is a key insight because in the multidimensional setting there are multiple cutoffs for each of the different biomarkers and the resulting space of possible combination rules lacks the ordering that is necessary to construct an ROC curve. The index $\alpha$ effectively provides an ordering of rules regardless of the number of markers, and thus a provides a unified definition of the ROC curve that applies to both the single and multiple marker settings.

RESULT 1 For each attainable false-positive rate, the corresponding point on the logic rule ROC curve maximizes the true-positive rate.

This observation can be proved by contradiction. For, suppose the point $(FP_o, TP_o)$ is the point on the logic rule ROC curve that corresponds to $\alpha$, and there exists another rule $R^*$ with false-positive rate $FP_1$ equal to $FP_o$, and a true-positive rate $TP_1$ that exceeds $TP_o$. Then, the weighted misclassification error rate for $R^*$ is even smaller than that associated with $(FP_o, TP_o)$, i.e. $\alpha FP_1 + FN_1 < \alpha FP_o + FN_o$, where $FN_1 = 1 - TP_1$ and $FN_o = 1 - TP_o$. However, if this is the case, then by Definition 1, $TP_1$

must be the value of the logic rule ROC curve at false-positive rate $FP_o$. Therefore, the point $(FP_o, TP_o)$ cannot be on the logic rule ROC curve.

The importance of this result is that it allows us to establish the connection between our method and previous approaches to developing and evaluating tests based on multiple biomarkers. For example, Baker (2000) discretizes a bivariate biomarker space and notes that with $n$ intervals in one dimension and $m$ intervals in another dimension, each split of the resulting $n$ by $m$ space in two corresponds to a classification rule. Plotting the true-positive versus the false-positive rates for each rule yields a cloud of points. Baker defines the ROC curve as the one connecting these points that lies 'highest and farthest to the left' and provided an algorithm from the econometric literature to derive it. Noting that the rules considered by Baker correspond to the set of logic rules given the assumed discretization, we observe that his algorithm identifies the logic rules that maximize the true-positive rate for each observed false-positive rate. By Result 1, these are theoretically the same as the optimal rules identified by the logic regression algorithm.

In practice, the results of our algorithm and that of Baker (2000) will likely differ because of our use of a probabilistic algorithm to search the rule space, our cross validation approach and our use of training and test data sets, all of which introduce some randomness into the procedure.

A similar argument illustrates the equivalence of our ROC concept and that of McIntosh and Pepe (2002). These authors show that the optimal combination test, namely the one that maximizes the true-positive rate for any specific false-positive rate, should be based on the risk score defined as the probability of disease given the observed biomarker values. There are many ways to model the probability of disease, but if a logic regression model is used, then the resulting space of possible combination rules corresponds to the space of logic rules, and their ROC curve for the optimal combination test will match ours.

A consequence of this result is that the logic rule ROC curve is monotone increasing. In practice, however, we consider a grid of values for $\alpha$, $\{\alpha_1, \alpha_2, \ldots \alpha_k\}$, and estimate the logic rule ROC curve as follows. First, the classification algorithm is run for each $\alpha$ value to select the corresponding logic rule. Then the true-positive rates for the selected logic rules are plotted against the associated false-positive rates. Because the classification algorithm may not always identify the optimal rule in practice, the estimated ROC curve may exhibit some non-monotonicity.

Note that each $\alpha$ value yields a corresponding point on the logic rule ROC curve. Therefore, the logic rule ROC curve is defined by a set of rules, $\mathbf{R}^* = \{R^*_{\alpha_j}; j = 1, \ldots K\}$, where the $\alpha$ are in decreasing order, and the corresponding rules are ordered by increasing false-positive rate. Each of these rules may be thought of as defining a subspace of the two-dimensional biomarker space, within which all subjects will be declared test-positive. This is analogous to the positivity region of Baker (2000). For a given rule, $R$, we say that an individual's test result is in $R$ if he would be declared positive on the basis of the rule $R$.

### 2.5 *Comparing diagnostic tests*

The area under the ROC curve (AUC) is a standard omnibus-type statistic for comparing diagnostic rules based on tests with continuous outcomes. In the one-dimensional setting where there is a single marker $X$, the AUC is interpretable as $P[X^D > X^{\bar{D}}]$, where $X^D$ and $X^{\bar{D}}$ are marker values for randomly selected diseased and healthy individuals respectively. Note that this can be rewritten as $P[X^D > r \text{ and } X^{\bar{D}} \leqslant r \text{ for some } r \in (-\infty, \infty)]$. Thus, the condition $X^D > X^{\bar{D}}$ is equivalent to the existence of a threshold, or one-dimensional rule, separating the two biomarker values, such that the test result for the diseased person is positive and that for the health person is negative. This notion of a separating rule generalizes directly to multiple dimensions and provides us with a concept that is analogous to the area under the curve.

DEFINITION 2 For a given logic rule ROC curve defined by set of rules $\mathbf{R}^*$, the probability of correct classification (PCC) is the probability that for any randomly selected pair of diseased and healthy individuals, with marker values $(X^D, Y^D)$ and $(X^{\bar{D}}, Y^{\bar{D}})$, there exists a rule $R$ in $\mathbf{R}^*$ such that $(X^D, Y^D)$ is in $R$ and $(X^{\bar{D}}, Y^{\bar{D}})$ is not.

By the preceding discussion, the PCC is equal to the AUC in the case of a single marker. In the case of multiple markers, however, the PCC is provably equal to the AUC only if the rules comprising $\mathbf{R}^*$ are nested.

DEFINITION 3 The rules in the set $\mathbf{R}^*$ are nested if the subspace defined by $R^*_{\alpha_i}$ is contained within that defined by $R^*_{\alpha_j}$ for all $i < j$.

As an example of nested and non-nested rules, consider the rules $R: \quad X > x \quad$ and $Y < 1$ and $R': \quad X > 2 \quad$ and $Y < 2$. If $x \geqslant 2$ then the rules are nested, otherwise they are not.

Although we anticipate that the subspaces defined by the rules $R^*_{\alpha_j}$ will generally increase in size as $j$ increases, they may not all nest, especially if we are attempting to constrain the rule size for purposes of predictive accuracy. In the case where the rule set is nested, we have the following result.

RESULT 2 For the logic rule ROC curve defined by $\mathbf{R}^* = \{R^*_{\alpha_j}; j = 1, \ldots K\}$, if the set $\mathbf{R}^*$ is nested, then the PCC is equal to the area under the logic rule ROC curve calculated by numerical integration.

PROOF 1 To prove this result, denote the sequence of rules in $\mathbf{R}^*$ by $R_1, R_2, \ldots R_K$ corresponding to increasing false-positive rates $FP_1, FP_2, \ldots FP_K$. By Definition 2, the PCC is given by $P[(X^D, Y^D) \in R$ and $(X^{\bar{D}}, Y^{\bar{D}}) \notin R$ for some $R \in \mathbf{R}^*]$. But this can be rewritten as

$$\text{PCC} = \sum_{m=1}^{K} P[(X^D, Y^D) \in R_m \text{ and } (X^{\bar{D}}, Y^{\bar{D}}) \notin R_m \text{ but } (X^{\bar{D}}, Y^{\bar{D}}) \in R_k \text{ for } k \geqslant m+1],$$

where $R_{K+1}$ classifies all subjects as positive. Equivalently,

$$\text{PCC} = \sum_{m=1}^{K} P[(X^D, Y^D) \in R_m] P[(X^{\bar{D}}, Y^{\bar{D}}) \notin R_m \text{ but } (X^{\bar{D}}, Y^{\bar{D}}) \in R_{m+1}].$$

But this is just equal to $\sum_{m=1}^{K} TP_m(FP_{m+1} - FP_m)$, where $TP_m$ is the logic rule ROC curve value corresponding to false-positive rate $FP_m$ and $FP_{K+1} = 1$. This is simply the area under the logic rule ROC curve calculated by numerical integration. □

Estimation of the PCC is straightforward. If the data for diseased individuals are denoted $(X^D_i, Y^D_i)$ and those for non-diseased individuals are $(X^{\bar{D}}_j, Y^{\bar{D}}_j)$, then

$$\widehat{\text{PCC}} = \sum_{i=1}^{n_D} \sum_{j=1}^{n_{\bar{D}}} I_{ij} / n_D n_{\bar{D}},$$

where $I_{ij} = I((X^D_i, Y^D_i) \in R$ and $(X^{\bar{D}}_j, Y^{\bar{D}}_j) \notin R$ for some $R \in \mathbf{R}^*)$.

The discretization of the biomarker ranges and the finite set of values considered for $\alpha$ may lead to ties between pairs of diseased and non-diseased observations in the sense that for every $R$ in $\mathbf{R}^*$, both observations will either be simultaneously in $R$ or not. We label these pairs of observations as neutral pairs. In practice, we assign $I_{ij}$ to one for half of the neutral pairs and to zero for the other half. When

the rules are nested, the corresponding estimate of the PCC corresponds to calculating the empirical AUC using the trapezoidal rule.

When the rules in $\mathbf{R}^*$ are not all nested, there will exist pairs $(X_i^D, Y_i^D)$ and $(X_j^{\bar{D}}, Y_j^{\bar{D}})$ such that $(X_i^D, Y_i^D) \in R$ and $(X_j^{\bar{D}}, Y_j^{\bar{D}}) \notin R$ but, in addition, $(X_j^{\bar{D}}, Y_j^{\bar{D}}) \in R'$ and $(X_i^D, Y_i^D) \notin R'$ for at least two rules $R$ and $R'$ in $\mathbf{R}^*$. We call these pairs of points *inconsistent* under $\mathbf{R}^*$ because although there exists a rule that separates the points and correctly classifies them as diseased and not diseased, there also exists a rule that separates the points and classifies them incorrectly. The relative frequency of inconsistent points is directly related to (a) the number of non-nested rules in the sequence of rules ordered by false-positive rate and (b) the probability content of the non-nested regions. With a low frequency of inconsistent points, the probability content of the non-nested regions will be low and the estimate of the PCC based on Definition 3 will approximate the AUC estimated by numerical integration. This is illustrated in our application.

Our definition of the PCC allows us to compare combination tests with competing tests using methods developed for AUC regression (Dodd and Pepe, 2002a). These methods allow us to model the AUC (or the PCC) as a function of covariates using generalized linear models. As an example, consider comparing the PCC for the test combining TPSA and RPSA with the PCC for TPSA alone. By Definition 2, the PCC for the combination test is simply the expectation of the binary variables $I_{ij}$, and the PCC for the TPSA-based test is also an expectation of binary variables given by $I(X_i^D > X_j^{\bar{D}})$; these expectations can be compared using binary regression with a single covariate representing test type (combination test or TPSA test). In this case, a set of indicators $\{I_{ij}\}$ is defined for each test type and the resulting $2 \times n_1 \times n_2$ indicator variables are considered as response variables in the analysis. This approach is also useful because it allows us to adjust the comparison for other covariates that might affect the result. For example, among prostate cancer cases in the Physicians' Health Study, the time from testing to diagnosis ranges from 1 to 12 years. We anticipate that sensitivity will depend strongly on this interval.

When computing variances of the regression parameter estimates for hypothesis testing purposes, it is important to recognize that the binary $I_{ij}$ variables are cross-correlated. Like Dodd and Pepe (2002a), we use bootstrapping to estimate parameter variances, and use the asymptotic normality of the parameter estimates for hypothesis testing. For each bootstrap sample, we obtain the logic rule ROC curve as described above, with one exception; for computational efficiency, we specify a model size of four, which we have found to produce similar results to those obtained when selecting a rule of size at most four by cross-validation. We then use binary regression to compare the PCC for the logic rule and the PCC for TPSA, adjusting for the time interval from testing to diagnosis for the disease cases.

## 3. THE PHYSICIANS' HEALTH STUDY

The Physicians' Health Study (PHS) was a randomized, placebo-controlled trial of aspirin and beta-carotene among 22 071 US physicians aged 40–84 years in 1982 (Gann *et al.*, 1995). At enrollment, 68% of participants provided a blood sample which was stored. Subsequently, serum from 430 men diagnosed with prostate cancer up to 12 years following enrollment was re-assayed for PSA and free PSA (Gann *et al.*, 2002). The majority of these cases were diagnosed prior to widespread adoption of the PSA test for prostate cancer screening. TPSA and RPSA data were available for these 430 cases and 1642 age-matched controls who had not been diagnosed with prostate cancer. For the combination rules, we discretized the TPSA range by cutpoints {1, 2, 3, 4, 5, 6, 7, 8, 9, 10}, where all measurements are in ng ml$^{-1}$ units. Cutpoints above 10 ng ml$^{-1}$ were not considered because values of TPSA in this range are generally considered sufficiently high to recommend biopsy in the absence of any additional information. Similarly, we discretized the RPSA range by cutpoints {0.05, 0.1, 0.15, 0.2, 0.25, 0.3, 0.35, 0.4, 0.45, 0.5}; based

Table 1. *Characteristics of cases and controls. Shown are data from cases and controls with both TPSA and RPSA values available at enrollment. Unless otherwise specified, numbers in parentheses are standard deviations*

| | Cases ($n = 429$) | Controls ($n = 1640$) |
|---|---|---|
| Mean age at test (years) | 60.29 (7.22) | 60.66 (7.33) |
| Mean TPSA at test (ng ml$^{-1}$) | 5.50 (11.08) | 1.84 (2.68) |
| Number with TPSA $> 4.0$ ng ml$^{-1}$ (%) | 147 (34.3) | 144 (8.8) |
| Number with TPSA $> 10.0$ ng ml$^{-1}$ (%) | 49 (11.4) | 21 (1.3) |
| Mean RPSA level | 0.20 (0.12) | 0.30 (0.15) |
| Number with RPSA $< 0.2$ (%) | 247 (57.6) | 403 (24.5) |
| Mean time from test to diagnosis (years) | 8.57 | na |

on clinical information, we felt that higher values of RPSA would not feature in the combination tests of interest.

The prostate cancer substudy of the Physicians' Health Study (Gann *et al.*, 1995) is possibly one of the most cited studies regarding the operating characteristics of TPSA. Since it is a retrospective, longitudinal, repository study (a Phase 3 study in the lexicon of Pepe *et al.* (2001)), it is not subject to the typical problems like selection bias and verification bias that are present in prospective screening studies (Begg, 1991). However, the definitions of sensitivity and specificity in this setting differ somewhat from the traditional definitions, namely the probability of a positive test given disease is present at the time of the test, and the probability of a negative test, given no disease present at that time. Rather, sensitivity here is the probability of a positive test given a *future* diagnosis of disease within a maximum time interval following the test, and similarly, specificity is the probability of a negative test given no future diagnosis of disease within this time. Thus, disease status is not ascertained for either cases or controls at the time of the test. Given that cases may be diagnosed up to 12 years after the time of their test, we anticipate that sensitivity for this group as a whole will be somewhat lower than it would be in the prospective setting and that sensitivity will likely depend on the time between testing and diagnosis.

## 4. RESULTS

Table 1 summarizes key characteristics of cases and controls. As noted by Gann *et al.* (2002), the age distribution at the time of the test was similar for cases and controls due to the age-matched design. PSA levels were significantly higher among cases ($p < 0.01$, Wilcoxon rank-sum test), as were complexed PSA levels ($p < 0.01$). The average ratio of free to total PSA was significantly lower among cases ($p < 0.01$). All of these differences were observed in spite of the median time from test to diagnosis for cases being 8 years. Figure 1 provides a scatterplot of the TPSA and RPSA results for the cases and controls in the study. For display purposes we have plotted data from a random 50% of controls, and have cut off the horizontal axis at TPSA $= 20$ ng ml$^{-1}$; 17 cases and 8 controls had TPSA values above 20 ng ml$^{-1}$. All of these cases and 6 of the 8 controls also had RPSA values below 0.2. The plot shows that a significant proportion of cases have TPSA values below the conventional cutoff of 4.0 ng ml$^{-1}$, and that a number of controls have TPSA values above this cutoff. However, the cases with TPSA below 4.0 ng ml$^{-1}$ tend to have longer time intervals between testing and diagnosis than those with PSA above 4.0 ng ml$^{-1}$ (7.5 versus 9.1 years on average, Wilcoxon rank-sum $p < 0.001$) and the controls with TPSA above 4.0 ng ml$^{-1}$ tend to be older than those with lower TPSA values (60.2 versus 65.1 on average, Wilcoxon rank-sum $p < 0.001$).

Figures 3 and 4 display the results of the logic regression to determine optimal combination rules corresponding to different relative weights for cases and controls. To avoid overfitting, we randomly

divided our data into two subsets, a training dataset consisting of two-thirds of the cases and the controls, and a test dataset consisting of the remaining one-third. We used the training dataset to identify logic rules for each value of $\alpha$ by cross-validation, but evaluated the logic rule ROC curve and PCC on the test data. This procedure (splitting the data two-thirds/one-third, identifying logic rules on the training data and estimating the ROC/PCC on the test data) was repeated 25 times, with different random splits of the whole dataset for each run. Results presented correspond to the run yielding values for the logic rule PCC and the TPSA-based AUC that were closest to their means over the 25 runs. This run was selected so as to provide results for what might be considered a 'typical' rather than an 'extreme' split of the data into test and training sets.

Figure 3 plots a sample of the rules themselves with varying $\alpha$ weights and corresponding false-positive and false-negative rates based on the complete data. This figure shows how, as the weight of false-positive relative to false-negative errors increases, the selected logic rule becomes more stringent, simultaneously reducing both true- and false-positive rates. The top two rules are nested, as are the bottom two rules, but the four rules together do not constitute a nested set. While the rules that were obtained all have the form of step-functions, the logic regression algorithm could have come up with any shape rules, including combinations of disjoint regions; the only restriction being on the size of the rule.

Figure 4 shows the ROC curves for the TPSA-based rule as well as for the combination rule. For fairness in the comparison, we evaluated a discretized TPSA-based rule with possible thresholds (in ng ml$^{-1}$) given by $(1, 1.5, 2, 2.5, 3, \ldots, 10)$; this discretization yielded a frequency of neutral rules that was similar to the sum of neutral and inconsistent rules for the logic rule ROC curve (approximately 8%). Since high specificity is important in cancer screening studies, the plots also show the ROC curve values for false-positive rates below 20%. The results indicate an apparent advantage for the logic rules within this region. Indeed, the plot shows that the classification algorithm identifies logic rules that have both lower false-positive and false-negative rates than the standard TPSA $>$ 4.0 ng ml$^{-2}$ rule, which has sensitivity equal to 33.6% and false-positive rate equal to 9.5%. As an example, the combination rule that most closely matches the sensitivity of the TPSA-based rule is (TPSA $>$ 1.0 ng ml$^{-1}$ AND RPSA $<$ 0.1) OR (TPSA $>$ 3.0 ng ml$^{-1}$ AND RPSA $<$ 0.15), which has sensitivity equal to 34.3% and a false-positive rate of 5.9%. The presence of these rules is consistent with the findings of Gann *et al.* (2002), that it is possible to identify combination rules with improved sensitivity and specificity relative to the standard TPSA $>$ 4.0 ng ml$^{-1}$ rule. However, it is important to test whether these apparent improvements are in fact statistically significant.

The area under the ROC curve for the TPSA-based rule is 0.747 and the PCC for the logic rule ROC curve is 0.752. The PCC for the logic rule closely approximates the area under the curve computed by numerical integration (trapezoidal rule), namely 0.749. The frequency of inconsistent pairs is low (3.1%) which explains the concordance; our estimate of the PCC assumes that separating rules in the sense of Definition 2 exist for half of these pairs.

For comparing tests, we used the following binary regression model:

$$\text{logit}(p_{ijk}) = \beta_0 + \beta_1 X_k + \beta_2 T_i + \beta_3 X_k T_i,$$

where $p_{ijk}$ is the probability that a rule of test type $k$ exists, separating case observation $i$ from control observation $j$, $X_k$ is an indicator of test type, and $T_i$ denotes time interval from test to diagnosis for case observation $i$. The $T_i$ term allows us to incorporate the impact of this time interval, while the interaction term allows us to discern whether the relative performance of the two tests changes with time prior to diagnosis and thus to determine whether one test might diagnose disease earlier in its natural history than the other (Etzioni *et al.*, 1999). In a prospective screening setting, this information would naturally not be useful since time prior to clinical diagnosis is not known at the time of screen detection. However, when making screening policy decisions, it is important to know whether one test is able to detect disease

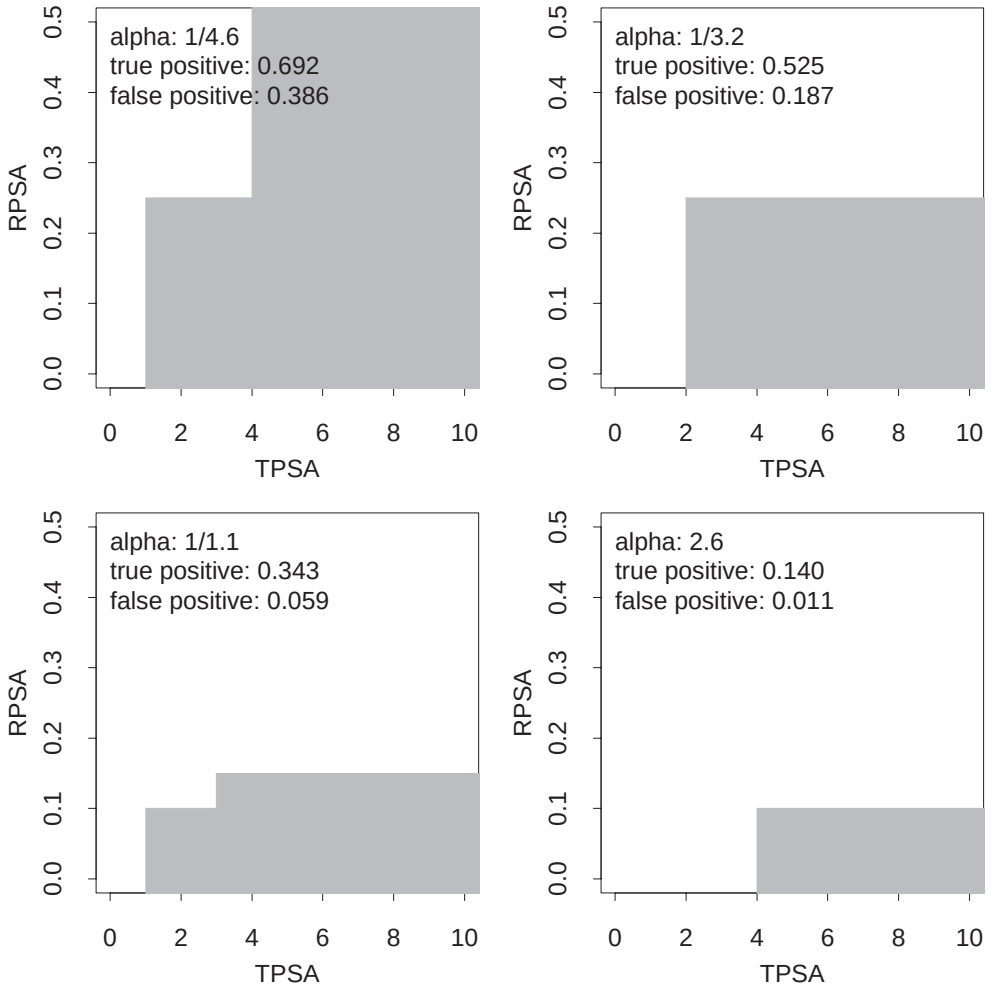


Fig. 3. Logic rules identified by the logic regression algorithm for different indexes $\alpha$ with corresponding true- and false-positive rates.

earlier in its natural history than others, since such a test could ultimately lead to improved effectiveness and cost-effectiveness.

Our estimate of the odds ratio corresponding to the coefficient $\beta_3$ was $-0.0523$ with bootstrap $Z$-value given by $-0.179$, indicating that the relative performance of the two tests did not change over time. Eliminating the interaction term, the coefficient estimate for $\beta_2$ was $-0.107$ with bootstrap $Z$-value given by $-2.181$, indicating that, as expected, diagnostic performance for both tests degrades as the time from testing to diagnosis increases. The results for $\beta_1$ (coefficient estimate 0.0697, $Z$-value 0.694) show a statistically non-significant improvement in performance associated with the combination test; this translates into only a modest improvement in the PCC as shown in Table 2. In addition, the lack of a significant interaction between the test type and time variables shows that the combination test does not appear to identify disease cases sooner than the test based solely on TPSA.

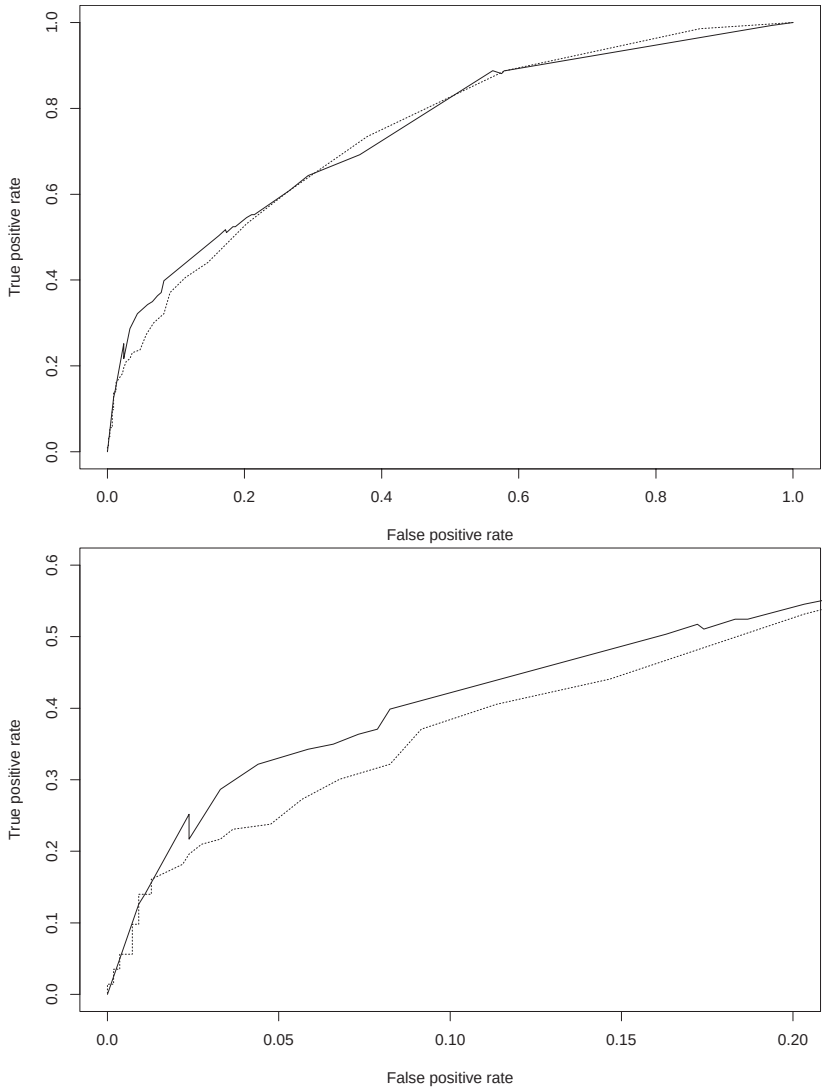


Fig. 4. Logic rule ROC curve for the combination of TPSA and RPSA and the rule based on TPSA. Data are split into training (2/3) and test (1/3) set. Logic regression is run on the training set and the rules identified are evaluated on the test set; this procedure is repeated 25 times. The ROC curves are plotted for the test data yielding values for the logic rule PCC and the TPSA-based AUC closest to their means over the 25 runs. (a) Entire curve. (b) Portion of the curve with false-positive rates ranging from 0 to 0.2 (partial ROC curve). Solid curve: combination test. Dashed curve: TPSA-based test. The lack of monotonicity of the ROC curve is likely due to our restriction of the rule space to rules of size 4 and/or the use of test data to evaluate and plot the ROC curves.

## 5. DISCUSSION

Current scientific advances promise that many novel biomarkers will soon become available for use in early detection and prognostication. Since it is difficult to find single biomarkers that perform well, attention has shifted to panels of biomarkers, and tests that combine marker values. However, standard statistical approaches for identifying and evaluating combination tests are not well developed. The

Table 2. *Fitted values of the PCC for the TPSA-based test and the logic rule combining TPSA and RPSA by time prior to diagnosis for cases. Data are split into training* $(2/3)$ *and test* $(1/3)$ *set. Logic regression is run on the training set and the rules identified are evaluated on the test set; this procedure is repeated 25 times. Results are based on the test data yielding values for the logic rule PCC and the TPSA-based AUC closest to their means over the 25 runs*

| Years from test to diagnosis | TPSA PCC | Logic Rule PCC |
|---|---|---|
| 1 | 0.890 | 0.886 |
| 2 | 0.877 | 0.874 |
| 3 | 0.863 | 0.859 |
| 4 | 0.843 | 0.843 |
| 5 | 0.831 | 0.825 |
| 6 | 0.812 | 0.806 |
| 7 | 0.792 | 0.786 |
| 8 | 0.770 | 0.763 |
| 9 | 0.747 | 0.740 |
| 10 | 0.721 | 0.714 |
| 11 | 0.695 | 0.687 |
| 12 | 0.667 | 0.659 |

RPSA/TPSA controversy illustrates the consequences of the lack of a formal methodology for quantifying the diagnostic gains associated with combining biomarkers. A large clinical literature exists concerning appropriate ways to use information on RPSA with TPSA measures, but results vary and conclusions are mixed. Although this is no doubt partly due to differences in study design and population characteristics, the studies also differ in their analytic approaches. Most analyses are exploratory and the vast majority are ad hoc from a statistical point of view. Typically, results pertaining to specific combination rules are presented, suggesting that improvements in false-positive rates can be attained with little or no decline in true-positive rates. However, it is not clear how the combination rules have been identified, nor whether the apparent improvements are statistically significant.

Statistical methods like ROC curve analysis have become more or less standard tools for evaluating diagnostic performance when test results can be reduced to a single measurement. However, it has not been clear how to extend these methods to the multidimensional setting of combination tests. Some methods have been developed to deal with linear combinations of markers (e.g. Pepe and Thompson, 2000), but methods for logic combinations, which are most clinically appealing, are largely lacking. An exception to this is the recent article by Baker (2000). The main problem when considering logic rules is that the rule space is multidimensional and unordered. Therefore, the ROC curve, which relies on an ordering of the rule space, is difficult to define. We have proposed a mechanism for constructing such an ordering that is intuitively reasonable, namely to order the rules according to the $\alpha$ value for which they are optimal. Rules that are never optimal for any $\alpha$ are dominated by other rules and need not be considered. The result is a unifying and interpretable concept of the ROC curve and the probability of correct classification which is analogous to the AUC, so that these measures can be used to quantify the discriminating capacity of tests based on any number of biomarkers. An additional advantage of our approach is that it is applicable to essentially any class of combination tests, and not solely the logic combinations, although we have focused on this class because of its clinical utility and interpretability.

In comparing the standard TPSA-based test with tests combining TPSA and RPSA, we found the PCC for the combination to be only slightly higher than that for the TPSA-based test, suggesting that at best a modest overall improvement in discrimination due to combining the biomarkers might be expected; this finding is consistent with the analysis of Baker (2000). With widespread use of the test, however, even a small improvement in diagnostic performance could translate into a clinically important reduction in the number of unnecessary prostate biopsies performed. The improvement could be greater when comparing specific TPSA thresholds with specific combination rules, when considering population subgroups (e.g. older men), or within specific regions of the ROC curve (e.g. for false-positive rates below a certain threshold) in which case partial ROC curves and their areas are of most interest (Baker, 2000; Dodd and Pepe, 2003). In the case of a single marker, $Y$, the partial AUC is the joint probability that $Y^D > Y^{\bar{D}}$ and that $Y^{\bar{D}}$ falls within a region of interest (i.e. below a specified quantile defined by a target false-positive rate). For logic combination tests, the partial AUC (or, more accurately, the PCC) would similarly be the joint probability that there exists a logic rule separating the two points and that the point for the non-diseased subject falls within the region of interest. However, it is not clear how to translate a specific quantile or false-positive rate into a 'region of interest' in the logic rule space. We plan to address this topic in future work.

When faced with an ROC curve which represents a collection of rules, it is natural to ask whether there is an 'optimal' or 'best' rule. Although the common wisdom is that rules 'higher and farther to the left' are preferable to rules that are 'lower and farther to the right' on the curve, the rule of choice will ultimately depend on the relative costs of false-positive and false-negative errors. The magnitudes of these costs will typically depend on the setting and even on the decision-maker. Discussion of these costs and their implications for rule selection are beyond the scope of this article; we have focused on a prerequisite to this step, namely whether, in the case of PSA testing, one should seek to determine an optimal TPSA-based test or, rather, one combining information on TPSA and RPSA. Baker (2000) presents one approach to identifying a 'target region' for the combination rule ROC curve which takes cost considerations into account. As noted by Baker (2000), in the case of screening interventions the target region should concentrate on rules with low false-positive rates. We have focused on methods for identifying clinically interpretable combination rules and have noted that rules combining TPSA and RPSA appear to improve diagnostic performance over the TPSA-based test in precisely this target region.

In their recent article that inspired this work, Gann *et al.* (2002) noted a 'need for methodological research to determine if sophisticated but user-friendly mathematical functions can provide better discrimination of cases and controls in various populations than our arbitrary selection of reflex ranges and cutpoints for testing.' The approach we have presented provides precisely such methodology, yielding a statistically valid and clinically accessible framework for quantifying the performance of rules that combine two or more biomarkers. We believe that these methods can circumvent the arbitrariness inherent in many prior studies of the role of free PSA in combination with total PSA. Moreover, they will provide a level of satisfaction that the space of clinically relevant rules has been systematically covered in the quest for combination rules with low false-positive and false-negative error rates.

## ACKNOWLEDGMENTS

Dr Etzioni's research was supported in part by R29 CA70227. Dr Kooperberg's research was supported in part by R29 CA74841 and P01 CA53996. Dr Pepe's research was supported in part by by R01 GM54438. We thank Dr Meir Stampfer and the investigators on the Physicians' Health Study for sharing the data from the PSA and prostate cancer substudy, which was supported by CA42182, CA58684 and CA57374.

Code for logic regression is available at `http://bear.fhcrc.org/~ingor/logic/`

## REFERENCES


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

BEDUSCHI, M. C. AND OESTERLING, J. E. (1998). Percent free prostate-specific antigen: the next frontier in prostate-specific antigen testing. *Urology* **51**, 98–109.

BEGG, C. B. (1991). Advances in statistical methodology for diagnostic medicine in the 1980's. *Statistics in Medicine* **10**, 1887–1995.

BRAWER, M. K. (2000). Prostate-specific antigen: Current status. *Ca: a Cancer Journal for Clinicians* **49**, 264–281.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. AND STONE, C. J. (1984). *Classification and Regression Trees*. Belmont, CA: Wadsworth.

CARLSON, G. D., CALVANESE, C. B. AND CHILDS, S. J. (1998). The appropriate lower limit for the percent free prostate-specific antigen reflex range. *Urology* **52**, 450–454.

CATALONA, W. J., SMITH, D. S., WOLFERT, R. L., WANG, T. J., RITTENHOUSE, H. G., RATLIFF, T. L. AND NADLER, R. B. (1995). Evaluation of percentage free serum prostate-specific antigen to improve specificity of prostate cancer screening. *Journal of the American Medical Association* **274**, 1214–1220.

CATALONA, W. J., SMITH, D. S. AND ORNSTEIN, D. K. (1997). Prostate cancer detection in men with serum PSA concentrations of 2.6 to 4.0 ng $ml^{-1}$ and benign prostate examination: enhancement of specificity with free PSA measurements. *Journal of the American Medical Association* **277**, 1452–1455.

DODD, L. AND PEPE, M. S. (2002a). *A Semi-parametric regression method for the area under the Receiver Operating Characteristic Curve*, Submitted for publication.

DODD, L. AND PEPE, M. S. (2003). *Partial AUC Estimation and Regression*, UW Biostatistics Working Paper Series. Working Paper 181. `http://www.bepress.com/uwbiostat/paper181/`.

ETZIONI, R., PEPE, M., LONGTON, G., HU, C. AND GOODMAN, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A prostate cancer case study. *Medical Decision Making* **19**, 242–251.

GANN, P. H., HANNEKENS, C. H. AND STAMPFER, M. J. (1995). A prospective evaluation of plasma prostate-specific antigen for detection of prostate cancer. *Journal of the American Medical Association* **74**, 298–294.

GANN, P. H., MA, J., CATALONA, W. J. AND STAMPFER, M. J. (2002). Strategies for Combining Total and Percent Free PSA for Detection of Prostate Cancer: a Prospective Evaluation. *Journal of Urology* **167**, 2427–2434.

MCINTOSH, M. W. AND PEPE, M. S. (2002). Combining several screening tests: Optimality of the risk score. *Biometrics* **58**, 657–664.

PARTIN, A. W., CATALONA, W. J., SOUTHWICK, P. C., SUBONG, E. N. P., GASIOR, G. H. AND CHAN, D. W. (1996). Analysis of percent free prostate-specific antigen (PSA) for prostate cancer detection: influence of total PSA, prostate volume and age. *Urology* **48**, 55–61.

PEPE, M. S., ETZIONI, R., FENG, Z., POTTER, J. D., THOMPSON, M. L., THRONQUIST, M., WINGET, M. AND YASUI, Y. (2001). Phases of biomarker development for early detection of cancer. *Journal of the National Cancer Institute* **93**, 1054–1061.

PEPE, M. S. AND THOMPSON, M. L. (2000). Combining diagnostic test results to increase accuracy. *Biostatistics* **1**, 123–140.

QUINLAN, J. R. (1993). *C4.5: Programs for Machine Learning*. San Mateo, CA: Morgan Kaufman.

REISSIGL, A., KLOCKER, H., PONTNER, J., FINK, K. *et al.* (1996). Usefulness of the ratio free/total prostate-specific antigen in addition to total PSA levels in prostate cancer screening. *Urology* **48**, 62–70.

RUCZINSKI, I., KOOPERBERG, C. AND LEBLANC, M. L. (2003). Logic regression. *Journal of Computational and Graphical Statistics* in press.

STENMAN, U. H., LEIONEN, J., ALFTHEN, H., RANNIKKO, S., TUHKANEN, K. AND ALFTHEN, O. (1991). A complex between prostate-specific antigen and alpha-1-antichymotrypsin is the major form of prostate-specific antigen in serum of patients with prostatic cancer: assay of the complex improves clinical sensitivity for cancer. *Cancer Research* **51**, 222–226.

VAN LAARHOVEN, P. J. AND AARTS, E. H. (1987). *Simulated Annealing: Theory and Applications*. Dordrecht: Kluwer.

[*Received April 22, 2002; first revision September 13, 2002; second revision December 18, 2002; accepted for publication December 20, 2002*]