STATISTICS IN MEDICINE
Statist. Med. 2005; **24**:1321–1338
Published online 29 November 2004 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/sim.2021

Identifying target populations for screening or not screening using logic regression

Holly Janes[1,*,†], Margaret Pepe[1,2], Charles Kooperberg[1,2] and Polly Newcomb[2]

[1]*Department of Biostatistics, University of Washington, Seattle, WA 98195, U.S.A.*
[2]*Division of Public Health Sciences, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, Seattle, WA 98195, U.S.A.*

SUMMARY

Colorectal cancer remains a significant public health concern despite the fact that effective screening procedures exist and that the disease is treatable when detected at early stages. Numerous risk factors for colon cancer have been identified, but none are very predictive alone. We sought to determine whether there are certain combinations of risk factors that distinguish well between cases and controls, and that could be used to identify subjects at particularly high or low risk of the disease to target screening. Using data from the Seattle site of the Colorectal Cancer Family Registry, we fit logic regression models to combine risk factor information. Logic regression is a methodology that identifies subsets of the population, described by Boolean combinations of binary coded risk factors. This method is well suited to situations in which interactions between many variables result in differences in disease risk. We found that neither the logic regression models nor stepwise logistic regression models fit for comparison resulted in criteria that could be used to direct subjects to screening. However, we believe that our novel statistical approach could be useful in settings where risk factors do discriminate between cases and controls, and illustrate this with a simulated data set. Copyright © 2004 John Wiley & Sons, Ltd.

KEY WORDS: logic regression; prediction; ROC curve; sensitivity; specificity; colon cancer

1. INTRODUCTION

Colorectal cancer is the third most common cancer in the United States. It is the second leading cause of cancer death among men and women despite the fact that the disease is treatable when detected at early stages [1] and that efficacious methods exist for early detection, namely sigmoidoscopy and colonoscopy [2, 3]. Such screening procedures can also detect

*Correspondence to: Holly Janes, Department of Biostatistics, University of Washington F-600 Health Sciences Building, Campus Mail Stop 357232, Seattle, WA 98195, U.S.A.
†E-mail: hjanes@u.washington.edu

Contract/grant sponsor: U0ICA 074794
Contract/grant sponsor: GM-54438
Contract/grant sponsor: CA 74841

Received May 2003
Accepted July 2004

pre-cancerous polyps that can then be removed, thus preventing disease from occurring. The key problem is that colon cancer screening is underutilized by the general public because it is invasive and costly, so that most disease is detected after it has progressed beyond the localized stage.

A range of risk factors for colon cancer have been identified. The motivation for the work described in this paper is to determine if subsets of the population with very high or low risk could be defined on the basis of these risk factors. This would provide an avenue for targeting screening efforts in the population. Individuals at high risk might be offered incentives or otherwise facilitated to undergo screening. Individuals at very low risk, on the other hand, might be allowed to forego screening and would not unnecessarily consume health care resources.

The Seattle site of the Colorectal Cancer Family Registry (CCFR) has collected data on colon cancer risk factors for 1680 cases and 1410 controls. This is a population-based case–control study with cases identified from the Puget Sound site of the Surveillance, Epidemiology, and End Results (SEER) registry, and controls, matched to cases on age and gender, selected at random from population lists [4]. As with most cancers, increasing age is the dominant risk factor for disease. Family history and male gender are also consistently associated with higher risk of disease. Other established risk factors include lack of physical exercise, intake of red meat, obesity (in males), alcohol and tobacco use. Use of aspirin and other non-steroidal anti-inflammatory agents, high intake of fruits and vegetables, folic acid taken as a food supplement and use of post-menopausal hormones have all been found to decrease the risk of colon cancer. Finally, a number of demographic and social factors have been linked with colon cancer (e.g. ethnicity and education) [5, 6].

Although epidemiologic associations exist with these factors, no one factor appears to be very predictive. Neither does a linear logistic model that combines risk factor information into a linear score appear to discriminate well between cases and controls (see Section 5.3). We suspected that interactions between multiple risk factors might be key in determining risk. For example, it might be that ('lack of exercise' or 'low dietary fibre') along with ('male gender' or 'female gender and not on post-menopausal hormones') would distinguish well between cases and controls. This subset is described by the logic tree shown in Figure 1. In this paper, we introduce logic regression as a method that could be useful for finding combinations of risk factors which discriminate between subjects at high and low risk of disease. Though a suitable criterion did not emerge from our risk factor information for colon cancer, we believe that

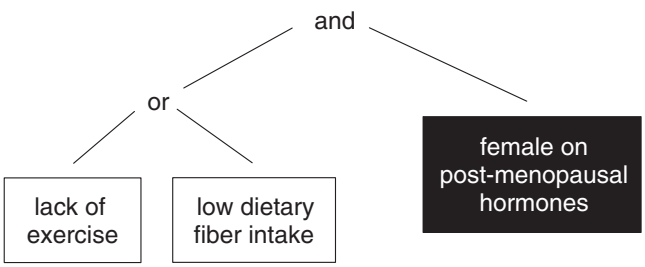

Figure 1. Example of a logic tree that evaluates to 1 if the Boolean expression illustrated is true. White letters on black denote negation of the entry.

logic regression is amenable to this task since it easily models high-order interactions between risk factors. In addition, a logic regression model yields a simple characterization of the subsets of the population at high risk, using logic trees, such as the tree in Figure 1. We begin in Section 2 with a description of logic regression. This is a new tree-based statistical technique for identifying subsets of the population defined by Boolean functions of binary coded risk factors, and is therefore well suited to our purposes. We contrast logic regression with another well known tree-based method for modelling binary data, classification and regression trees (CART), in Section 2. In Section 3, we describe the CCFR study in some detail. Section 4 is concerned with the evaluation of a fitted logic regression model for the purposes of developing criteria that could be used to direct subjects to screening sigmoidoscopy. Our results for colon cancer, described in Section 5 are disappointing in that useful criteria do not seem to emerge from the data. Nevertheless, we believe that the novel statistical approach we took could be useful in settings where interactions between risk factors do discriminate cases from controls. In Section 6, we demonstrate this with a simulated data set. We conclude in Section 7 with a discussion of the potential for pre-screening with risk factor information in health care and further refinement to the logic regression methodology that may facilitate its use for identifying pre-screening criteria.

2. LOGIC REGRESSION

Logic regression can be applied to any type of regression outcome as long as the proper scoring function is specified. We have a binary outcome and use deviance of logistic regression as the score function. For a given set of Boolean expressions, an example of which was given in Section 1, the logic regression model is a logistic regression model with those Boolean expressions as covariates. Specifically, we denote a Boolean expression with the binary variable L, where $L=1$ is 'true' and $L=0$ is 'false'. The model is written as

$$\operatorname{logit} P(D=1 \mid L_1,\ldots,L_P)=\alpha_0+\beta_1 L_1+\cdots+\beta_P L_P \tag{1}$$

What distinguishes logic regression from simple logistic regression with binary covariates is that the fitting algorithm both defines covariates for the model (using risk factor data) and estimates the regression coefficients simultaneously. The output from logic regression is represented as a series of trees, one for each Boolean predictor, L, and the associated regression coefficient. The logic tree for the expression defined earlier is shown in Figure 1.

Ruczinski *et al.* [7] provide a detailed description of logic regression and the simulated annealing algorithm used to fit it. They also contrast logic regression with other methods for modelling binary response data. Software for fitting logic regression models using the simulated annealing algorithm is available from http://www.bear.fhcrc.org/~ingor/logic.

Logic regression was proposed for settings where interactions between many variables give rise to large differences in response. This occurs, for example, in single nucleotide polymorphism association studies, where multiple genetic point mutations may be jointly associated with a disease outcome. See Reference [8] for a successful application of logic regression in this setting. We suspect that disease risk factors may behave similarly. Etzioni *et al.* [9] use logic regression to combine two prostate cancer biomarkers together. They use continuous biomarker data by defining multiple dichotomous predictors using various thresholds for the biomarkers. Ruczinski *et al.* [7] provide further examples of applications of logic regression.

CART is another tree-based method for modelling binary data [10]. The classification rule is displayed as a tree whose leaves are the two classes of interest (e.g. diseased and non-diseased), and whose branches correspond to dichotomized covariates. Each leaf is reached by one or more paths through the tree; to reach the leaf, all conditions along the path must be satisfied. Thus, a classification tree can be thought of as the collection of all paths that reach a leaf predicting class 1. Therefore, any classification tree can be written as a Boolean combination of covariates, as can a logic regression tree. (In the computer science literature, such rules are said to be in disjunctive normal form (DNF).) However, there are some Boolean expressions which can be very simply represented as logic trees, but which require fairly complicated classification trees [7]. It is the simplicity of logic trees which we hope to exploit in order to produce easily interpretable characterizations of high risk individuals.

In addition to the specification of the scoring function, the fitting algorithm for logic regression also requires specification of the number of logic trees (P in equation (1)) and the maximum number of variables, or leaves, that can make up a tree (three in the example in Figure 1). As with any adaptive regression methodology, larger models (those with more trees and leaves) typically fit better than smaller models. In this paper we chose model sizes *a priori*; for interpretability we fit models with four leaves per tree. More generally, one can select the size of the model with the data using techniques such as cross-validation or randomization tests, as described by Ruczinski *et al.* [7].

For a given model size, the selection of the best logic trees L_j is a non-trivial optimization problem. The logic regression algorithm that we implemented employs a simulated annealing algorithm. Simulated annealing [11] is a stochastic optimization algorithm similar to the Metropolis–Hastings algorithm for Markov chain Monte Carlo [7]. As with any stochastic optimization algorithm, there is no guarantee that the 'best' model is found, though with proper adjustment of various tuning parameters we can be confident that we have selected a good model.

3. THE REGISTRY DATA

The Seattle Familial Registry for Colorectal Cancer is a member of the International Colon Cancer Family Registry (CCFR). It was established in 1998 as a resource for studying the genetic epidemiology of colorectal cancer. From 1998 to 2002, cases aged 20–74 years of both genders diagnosed with incident colon or rectal cancer were identified from the Puget Sound SEER registry. Controls were randomly selected from two sampling frames. For cases age 20–64 years, controls were identified from lists of licensed drivers; for those age 65–74 years, controls were selected from files of the Health Care Financing Administration. All subjects completed an interviewer administered questionnaire on family and medical history, environmental and lifestyle factors, and screening history, and biological samples were collected [12]. Response rates were high (80 per cent for cases, 71 per cent for controls) [4].

The data used in this analysis are a subset of the registry data. We began with 769 cases and 657 controls, recruited in the last study year. We set aside one third of the cases and one third of the controls, randomly selected within age strata, for validation testing of the model.

Logic regression requires binary predictor variables, so we recoded variables into binary forms. Categorical covariates were coded as a set of indicator variables for each level of the covariate. Continuous covariates were coded as a series of threshold indicators. For example,

pack-years of smoking was coded as three indicators: (pack-years >0), (pack-years >9), and (pack-years >19). Where possible, thresholds were chosen to be quintiles of the covariate in the control population (with the exception of pack-years, for which thresholds were chosen *a priori*). Thresholds for BMI and height were chosen separately for men and women; the thresholds correspond to quintiles of the gender-specific control populations. Subjects who had a sigmoidoscopy more than 1 year prior to study enrollment were considered to have a screening history. For two covariates with a large amount of missingness (hours of physical exercise and fried poultry consumption), indicators of missingness were also included.

The data used to fit the logic regression model include 66 binary covariates. Since the logic regression algorithm currently cannot handle missing data, subjects with any missing covariates were not included in the analysis. Missingness was as large as 2.4 per cent for a given predictor. A total of 463 cases and 415 controls were used to fit the model.

4. OPERATING CHARACTERISTICS OF THE FITTED MODEL

4.1. The receiver operating characteristic (ROC) curve

Recall that the overall objective is to define criteria for who should or should not be recommended for clinical screening. We evaluate the sensitivity (true positive fraction (TPF)) and specificity (1—false positive fraction (FPF)) of criteria based on the risk factor model. Since the data are from a case–control study, with sampling dependent on disease status, we cannot evaluate predictive values directly from the data, but we can evaluate true and false positive fractions. It is natural to consider positivity criteria based on the risk score, $P(D=1\,|\,L_1,\ldots,L_P)$, or equivalently the linear predictor, exceeding a threshold c:

$$\text{`positive'} = \text{`}\beta_1 L_1 + \cdots + \beta_P L_P > c\text{'}$$

Such decision criteria are known to be optimal [13]. The associated true and false positive fractions,

$$\text{TPF}(c) = P(\text{positive}\,|\,\text{diseased})$$

and

$$\text{FPF}(c) = P(\text{positive}\,|\,\text{not diseased})$$

are quantities derived from cases and controls, respectively. A plot of $(\text{FPF}(c), \text{TPF}(c))$ displays the range of operating characteristics attainable with the risk factors. This plot is known as the ROC curve.

For our settings, we seek criteria which are either very sensitive and at least moderately specific, or very specific and at least moderately sensitive. If a very sensitive criterion were developed, we could be confident that we would not miss many cases by recommending that subjects who do not meet the criterion forego screening. This would give rise to a savings in health care resources. If a very specific criterion were presented, on the other hand, one might encourage subjects satisfying the criterion to avail of screening procedures, since these subjects are at relatively high risk of disease. We therefore focus on points on the ROC curve that relate either to high values for TPF or to small values for FPF.

4.2. Predictive values

The predictive values of a criterion quantify the risk of disease for subjects that are positive or negative on the criterion. These entities relate directly to the usefulness of the criterion in the population. However, they depend on disease prevalence, which cannot be determined from a case–control study. With our data, we can only obtain estimates of the true and false positive fractions associated with a criterion. These are the probabilities of criterion positivity given incident disease status, and we assume they are valid nationally. We then used the national SEER incidence rates for colorectal cancer (denoted by ρ) to calculate predictive values (PV), using the following relationships:

$$\begin{aligned}\text{Positive PV} &= P(D=1 \mid \text{positive}) \\ &= \rho\text{TPF}/\{\rho\text{TPF} + (1-\rho)\text{FPF}\} \\ \text{Negative PV} &= P(D=0 \mid \text{negative}) \\ &= (1-\rho)(1-\text{FPF})/\{(1-\rho)(1-\text{FPF}) + \rho(1-\text{TPF})\}\end{aligned}$$

Again, a criterion with a high positive PV could be useful for selecting subjects for clinical screening. Negative predictive values are always high for a rare disease and so tend to be less useful. However, it will be important to determine the proportion of the population that satisfy the criterion $\tau = \text{Prob(positive)}$, in order to assess the impact of using such a criterion in the population. We calculate τ with the formula:

$$\tau = \rho\text{TPF} + (1-\rho)\text{FPF}$$

4.3. Stratum-specific performance

As is typical of many case–control studies, the CCFR is designed so that controls are frequency matched with cases. Matching on gender and age (by decade) was implemented to control for these major confounders. The implications of matching are threefold: (i) the effects of age and gender on disease risk cannot be estimated. They are fixed in the sample by design; (ii) the effects of other risk factors can be estimated, but only within subpopulations defined by age and gender; (iii) and to do this, it is necessary to include age and gender as covariates in the model for disease risk [14]. We categorized age into five categories, which along with gender defines ten strata. A stratum-specific intercept, α_s for $s = 1, \ldots, 10$ was included in the model

$$\text{logit}\, P(D=1 \mid \text{age, gender, risk factors}) = \alpha_s + \beta_1 L_1 + \cdots + \beta_P L_P$$

The matching variables are included among the risk factors for defining the Boolean covariates in the model, since their interactions with other risk factors are estimable. If such occurs, the interpretation is that the relevant risk factor combinations or their effects differ amongst the strata.

Since the intercepts of our model, α_s, are biased due to the matching, we cannot assess the performance of the model as a predictor in the whole sample. Within matching strata, however, the intercepts are merely constants, so we can assess criteria such as '$\beta_1 L_1 + \cdots + \beta_P L_P > c$'

within strata. We therefore calculate the $(\mathrm{FPF}(c), \mathrm{TPF}(c))$ values using the cases and controls within each stratum. We also calculate predictive values, using stratum-specific incidence rates (available from SEER). Since it is not clear how best to summarize these operating characteristics across strata, particularly if they vary amongst strata, we report all stratum-specific values here.

5. RESULTS FOR COLON CANCER DATA

5.1. The simple one-tree model

We first fit a model with a single Boolean tree predictor, i.e. $P = 1$. The tree is shown in Figure 2. The odds ratio and 95 per cent confidence interval associated with the tree are $\exp(\hat{\beta}_1) = 2.9$ and $(2.1, 3.9)$, respectively, with p-value < 0.001.

The factors identified in the data concur with previous reports in the literature. Family history of disease and overweight (in males) are well established as colon cancer risk factors [5]. Less education is likely to be a surrogate for less healthy lifestyle and less access to health care resources amongst other things. It too has been found to be associated with higher risk of colon cancer. Women taking estrogen post-menopausally have a reduced risk of colon cancer. The logic tree indicates that having a family history of colon cancer or having less education defines a group at substantially increased risk of colon cancer. However, post-menopausal females in this group who take estrogen are not at increased risk unless they are substantially overweight. As a group, those satisfying the logic tree are estimated as having a relative risk of almost 3 compared to subjects of the same age and gender who do not satisfy the tree. This is likely an overestimate since it is estimated from the same data that selected this covariate on the basis of its association with risk in this data. We therefore re-estimated the relative risk associated with the tree using the validation data that we had set aside. The estimated age and gender adjusted relative risk is 3.0 (95 per cent confidence interval $= (2.0, 4.5)$, p-value < 0.001). The odds ratio estimate is the same as that based on the training data, although the confidence interval is wider because of the smaller sample size in the validation set.

With only one tree, the operating characteristics of the fitted model are very simple. There is only one distinct non-degenerate positivity criterion to consider, namely, whether or not the

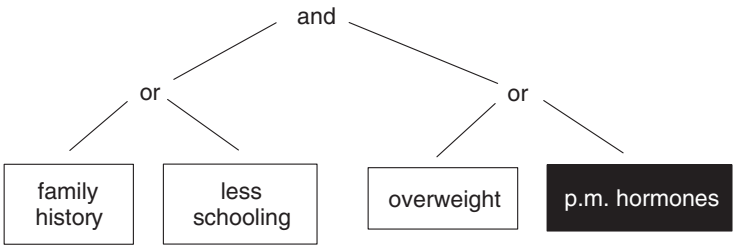

Figure 2. The single tree, L_1, fitted to the colon cancer data. The risk factors included are: family history (yes/no); less schooling (high school education or less); overweight (body mass index $> 26.6\,\mathrm{kg/m^2}$ for females, $> 27.3\,\mathrm{kg/m^2}$ for males); and p.m. hormones (women post-menopause ever taking hormones for more than 6 months).

tree is satisfied ($L_1 = 1$). The estimated sensitivity and specificity values for this criterion are shown for the eight strata that had >20 cases and controls (Table I). Again, we note that performance is similar with the validation and training data sets, although, as expected, there is more statistical variability with the smaller validation set. The sensitivities, averaging about 45–50 per cent, are not very high. We certainly could not use this criterion and consider screening to be unnecessary in the subpopulation that is criterion-negative because about half of diseased subjects are criterion negative. The specificity is better, averaging about 76 per cent across the strata. However, it may not be appropriate to use this criterion for targeting intense screening encouragement efforts either: about 24 per cent of non-diseased subjects would be unnecessarily enticed to undergo clinical screening with this criterion.

It is interesting that the tree, L_1, defines a group with a high relative risk of disease but does not yield a criterion with good operating characteristics. We show the stratum-specific odds ratios associated with L_1 in Table I, which are reasonably well summarized by the overall odds ratio $\exp(1.06) = 2.9$ from the fitted model. The odds ratios can be calculated directly from the sensitivity and specificity values as:

$$\text{Odds ratio} = \frac{\text{TPF}}{(1 - \text{TPF})} \frac{1 - \text{FPF}}{\text{FPF}} \tag{2}$$

From equation (2) we see that the odds ratio is a composite of the sensitivity and specificity. Clearly it will be large if either the sensitivity is large or if the specificity is large, since these yield small denominators, $(1 - \text{TPF})$ and FPF, respectively. However, it is notable that criteria with moderate sensitivity and specificity values can also have large odds ratios (Figure 3). This reinforces the need to examine the two components of the odds ratio, (FPF, TPF), not just their composite, for the sorts of applications we have in mind [15].

We now turn to the population performance of the criterion. Table I displays τ, the fractions of the population that are estimated to satisfy the criterion (the fraction for whom $L_1 = 1$). It ranges from 29 to 46 per cent across the strata. Note that the incidence of colon cancer is very low, ranging from about 20/100 000/year in 40–50 year old women to 364/100 000/year in 70–79 year old men [1]. This, along with the moderate specificity of the criterion, gives rise to low positive predictive values (Table I). The highest value is seen in 70–79 year old females where the incidence of colon cancer is estimated to be 8.1/1000 in women who are criterion positive. This seems unlikely to provide strong motivation for campaigning for screening in this population.

Recall that we chose *a priori* to have a model with four leaves. We performed a cross-validation analysis to assess whether our choice of model overfit the data. We found that the four-leaf model had a slightly higher cross-validated deviance than smaller models (a difference of less than 10 on a deviance scale), but we do not expect that this difference would be associated with meaningful differences in operating characteristics.

5.2. More subpopulations

We next fit models with two trees, $P = 2$. The model was fit six times, resulting in five unique models. Since the simulated annealing algorithm used to fit the logic regression models is not guaranteed to find the 'best' model, this variation is to be expected. On any given run, the model selected may correspond to a peak in the likelihood, but fitting the model several times allows us to determine if there is some model with an exceptionally good score. The five

Table I. Operating characteristics for the single tree model (Figure 2) for the colon cancer data.

Age (years)	Gender	Number of cases	Sensitivity (per cent)	Number of controls	Specificity (per cent)	Odds ratio	Criterion positive (τ) (per cent)	1 year positive PV (per cent)
(A) *Training data set*								
40–49	Female	26	46.15 (26.59, 66.63)	29	72.41 (52.76, 87.27)	2.25 (0.73, 6.91)	36.36	0.03
40–49	Male	21	47.62 (25.71, 70.22)	26	80.77 (60.65, 93.45)	3.82 (1.04, 13.98)	31.91	0.05
50–59	Female	74	43.24 (31.77, 55.28)	86	82.56 (72.87, 89.90)	3.61 (1.75, 7.43)	29.38	0.14
50–59	Male	73	41.10 (29.71, 53.23)	45	82.22 (67.95, 92.00)	3.23 (1.32, 7.90)	32.20	0.20
60–69	Female	91	45.05 (34.60, 55.84)	89	74.16 (63.79, 82.86)	2.35 (1.25, 4.41)	35.56	0.24
60–69	Male	95	45.26 (35.02, 55.81)	45	73.33 (58.05, 85.40)	2.27 (1.05, 4.93)	39.29	0.34
70–79	Female	61	59.02 (45.68, 71.45)	66	75.76 (63.64, 85.46)	4.50 (2.10, 9.62)	40.94	0.64
70–79	Male	44	50.00 (34.56, 65.44)	43	69.77 (53.88, 82.82)	2.31 (0.96, 5.56)	40.23	0.60
(B) *Validation data set*								
40–49	Female	16	50.00 (24.65, 75.35)	19	78.95 (54.44, 93.95)	3.75 (0.86, 16.40)	34.29	0.05
40–49	Male	12	25.00 (5.49, 57.19)	12	66.67 (34.89, 90.08)	0.67 (0.11, 3.93)	29.17	0.01
50–59	Female	35	40.00 (23.87, 57.89)	38	71.05 (54.10, 84.58)	1.64 (0.62, 4.33)	34.25	0.08
50–59	Male	26	30.77 (14.33, 51.79)	15	86.67 (59.54, 98.34)	2.89 (0.52, 15.91)	24.39	0.19
60–69	Female	54	48.15 (34.34, 62.16)	44	81.82 (67.29, 91.81)	4.18 (1.64, 10.63)	34.69	0.36
60–69	Male	47	44.68 (30.17, 59.88)	19	84.21 (60.42, 96.62)	4.31 (1.10, 16.79)	36.36	0.56
70–79	Female	46	56.52 (41.11, 71.07)	38	81.58 (65.67, 92.26)	5.76 (2.10, 15.75)	39.29	0.81
70–79	Male	19	52.63 (28.86, 75.55)	16	62.50 (35.43, 84.80)	1.85 (0.48, 7.18)	45.71	0.51

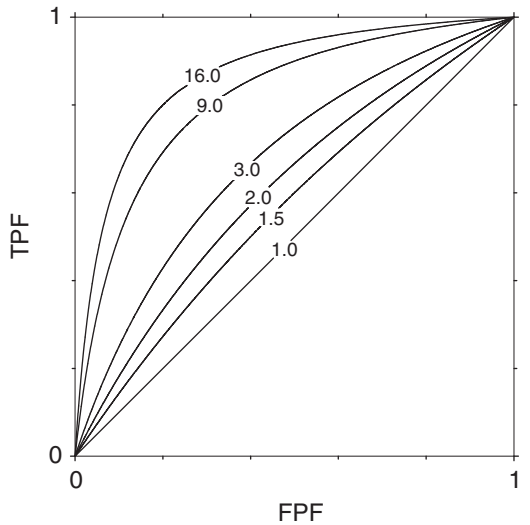

Figure 3. Contour plots for the odds ratio. (FPF, TPF) combinations that yield equal values for the odds ratio are connected. Shown are contours for odds ratios of 1.0, 1.5, 2.0, 3.0, 9.0, and 16.0.

models we found all had very similar scores, indicating that for this problem there are many models that perform equally well. We present the results for the model whose covariates we felt are most easily defined.

The two-tree model is shown in Figure 4. Interestingly the first tree, L_1, is the same as that arrived at when we allowed only one tree in the model. The estimated odds ratio, $3.0 = \exp(1.096)$, is also similar. The second tree, L_2, involves different risk factors, including one (poultry consumption) that has not been previously consistently implicated in colon cancer. The model with linear predictor $\beta_1 L_1 + \beta_2 L_2 = 1.096 L_1 + 0.777 L_2$ gives rise to three distinct non-degenerate criteria for defining subpopulations. Let us consider the operating characteristics for this model. The most specific criterion based on the model is where both trees are positive, which corresponds to choosing $c > 1.096 + 0.777$. The most sensitive non-trivial rule is where tree 1 or tree 2 is positive $c > 0.777$. The associated operating characteristics in the validation data are shown in Figure 5. The most specific criterion had an estimated specificity that averaged 89 per cent across strata, with corresponding average sensitivities of 25 per cent. If these numbers are accurate, it appears that 25 per cent of cases could be identified for screening with the criterion without referring more than 11 per cent of non-diseased subjects for unnecessary screening. The most sensitive criterion averaged 83 per cent with specificities that average 33 per cent across the strata. If these numbers are accurate, we could save 33 per cent of controls from unnecessary screening while continuing to screen the majority of cases. These operating characteristics are disappointing. We felt that neither the most sensitive nor the most specific rule would be useful in advising individuals to take advantage of or to forego screening.

As more trees are added to the model, this creates a broader range of criteria that can be investigated. There are, in fact, 2^P criteria that are formed from the linear predictor $\beta_1 L_1 + \cdots + \beta_P L_P$. This follows from the fact that the P binary logic trees partition the population

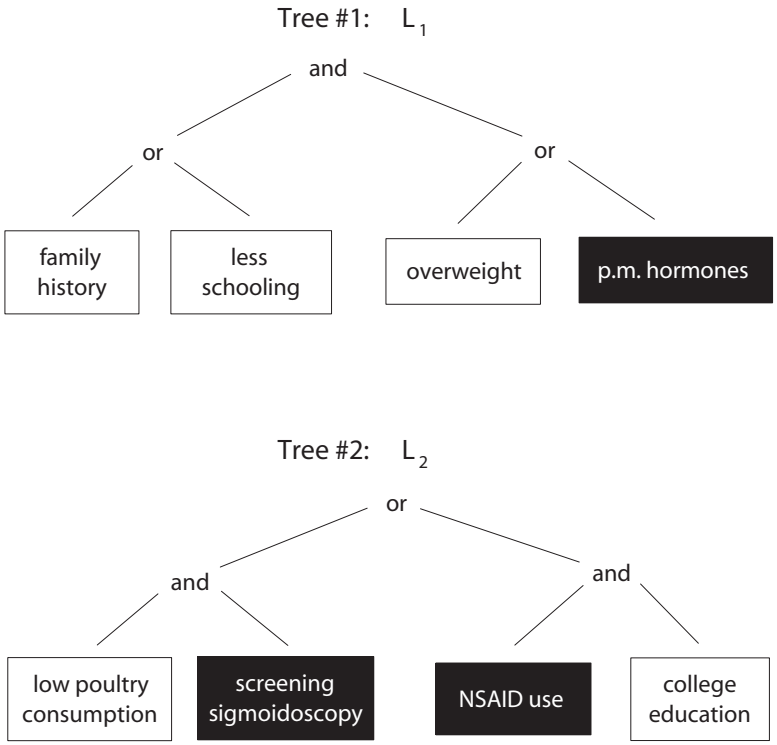

Figure 4. The trees L_1 (upper panel) and L_2 (lower panel) fit to the colon cancer data. The fitted age and gender adjusted model is $\beta_1 L_1 + \beta_2 L_2 = 1.096 L_1 + 0.777 L_2$. Variables in L_1 are described in Figure 1. Variables in L_2 are: low poultry consumption ($\leqslant 2$ servings per week); screening sigmoidoscopy (>1 year before study entry); NSAID use (>0.25 months using non-steroidal anti-inflammatory drugs); and college education (some college education).

into 2^P subgroups. In general, the operating characteristics associated with this model are represented as 2^P points along an ROC curve. We did not explore $P>2$, but this could be done in other applications.

5.3. *Comparison with linear logistic regression*

We fit a linear logistic model to the CCFR data. A stepwise algorithm yielded the results shown in Table II. Covariates whose statistical significance was $p<0.2$ were sequentially added to the null model. The operating characteristics for criteria based on this model are the (FPF, TPF) points corresponding to the rules

$$\gamma_1 X_1 + \gamma_2 X_2 + \cdots + \gamma_K X_K > c \tag{3}$$

where X_k denotes a covariate in the model, γ_k is the associated log odds ratio, and c is the threshold for the rule. Because of the large number of covariates, $K=9$, and the fact that some covariates are on a continuous scale, the (FPF, TPF) points map out a continuous ROC

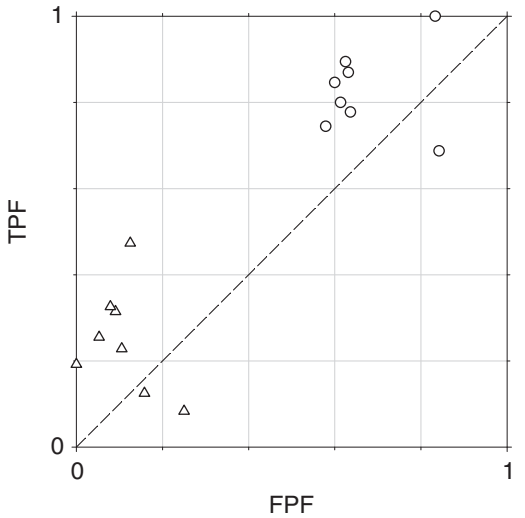

Figure 5. Operating characteristics for criteria based on the two-tree model. Each point represents a stratum with numbers of cases and controls shown in Table IB. Values for most sensitive (○) and most specific (△) criteria are displayed.

Table II. Exponentiated coefficients from a linear logistic regression model fit to the colon cancer data.

	Odds ratio	95 per cent CI
Education		
High school or less	1.00	
Some college	0.62	(0.43, 0.91)
College graduate	0.47	(0.32, 0.69)
Body mass index (kg/m^2)	1.04	(1.01, 1.07)
Calcium		
Months of use	0.96	(0.93, 0.99)
Family history of colon cancer		
Yes versus no	2.78	(1.81, 4.28)
Screening sigmoidoscopy		
Yes versus no	0.59	(0.40, 0.86)
Fried poultry		
Servings per week	1.04	(0.99, 1.10)
Poultry		
Servings per week	0.90	(0.82, 0.99)

Age and gender were included in the model.

curve for $c \in (-\infty, \infty)$. The curves may well vary across strata. We estimated stratum-specific ROC curves using the binormal model

$$\mathrm{ROC}(t) = \Phi(a_s + b_s \Phi^{-1}(t))$$

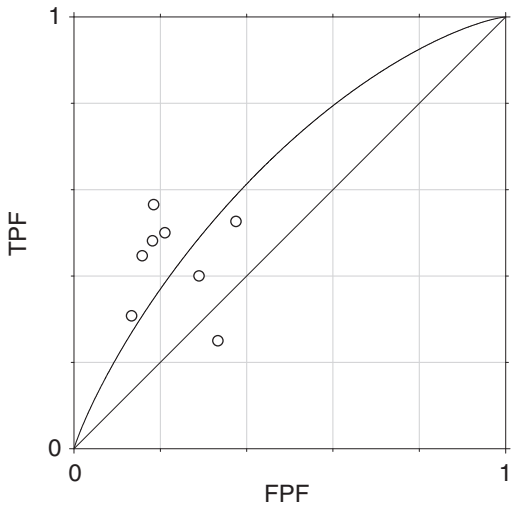

Figure 6. Operating characteristics associated with the linear logistic model. The average ROC curve is shown, $\mathrm{ROC}(t) = \Phi(\bar{a} + \bar{b}\Phi^{-1}(t))$ with $\bar{a} = 0.55$ and $\bar{b} = 1.05$. Estimated (FPF, TPF) points for the single tree logic regression model are also shown.

where Φ denotes the cumulative standard normal distribution function and (a_s, b_s) are stratum-specific ROC intercept and slope parameters. The LABROC algorithm was used to find parameter estimates [16]. The average curve

$$\mathrm{ROC}(t) = \Phi(\bar{a} + \bar{b}\Phi^{-1}(t))$$

where $\bar{a} = \sum_{s=1}^{S} a_s/S$ and $\bar{b} = \sum_{s=1}^{S} b_s/S$, is shown as the curve in Figure 6. Both the ROC curve and the (FPF, TPF) points associated with the logic regression model shown pertain to the validation data. As with the logic regression models, the risk factors do not yield criteria with adequate operating characteristics from the fitted linear logistic regression model.

In general, we prefer logic regression over linear logistic regression. A logistic regression model does not yield a simple characterization of the subset of the population at high risk. The subgroup is simply those subjects whose weighted average of risk factors, (3), is above a specified threshold. The logic trees, on the other hand, simply characterize the subset of the population that is at high risk, although this comes at a cost of some constraints on risk factor parametrization. Logic regression also easily models high-order interactions, while stepwise logistic regression does not. Though all possible interactions could be coded by hand and entered into a stepwise procedure, much modification would be needed to ensure that interactions not be included without their associated main effects. In this data set, however, there do not seem to be identifiable subsets of the population that are at risk, and both approaches yield inadequate prescreening criteria for colon cancer.

6. ILLUSTRATION WITH SIMULATED DATA SET

In order to validate the use of logic regression in a setting in which high-order combinations of covariates are important for predicting disease, we simulated such a data set. We assumed that disease risk is a function of categorized continuous covariates, since modelling the covariates continuously would have necessitated assuming an arbitrary functional form for the association. We generated a population with an age- and gender-specific covariate distribution similar to the controls in the colon cancer registry data. We set the size of the simulated population at $N = 7000$. Subjects in this hypothetical simulated population were at high risk for colon cancer if they were heavy males (BMI > 25.7) with a family history of colon cancer, or female smokers (pack-years > 0) who were not heavy (BMI $\leqslant 24.2$). This logic tree is shown in Figure 7. Those satisfying these conditions became cases in the simulation with probability 0.75, while those not in this subgroup became cases with probability 0.2. We then selected 100 cases and 100 controls at random from each of the 10 age and gender strata. The stratum-specific operating characteristics of the logic tree used to generate the data are contained in Table III. The fact that membership in the high risk subgroup is rare and that the large number of subjects outside of this group developed cancer by some other cause with probability 0.2 means that there are a large number of cases who are not described by the logic tree. Consequently, some of the stratum-specific sensitivities are very low (0–2 per cent). The specificities are high, a result of the rarity of the high risk subgroup (87–100 per cent).

A logic regression model with one tree and eight leaves, including age and gender effects, was fit to the simulated data (see Figure 8). By comparing Figures 7 and 8, we can see that the fitted tree is not exactly the same as the tree used to generate the data, but the high risk subgroups described are very similar. In fact, only 15 of the total 2000 subjects are differentially classified by the two trees. It is possible that further model selection would result in a model that is even more similar to the true model. For comparison, a stepwise logistic regression model, also including age and gender, was fit to the data. The operating characteristic of the logic and logistic models were assessed using a very large validation data set ($N = 78\,000$). The stratum-specific empirical ROC curves for the logistic model are shown

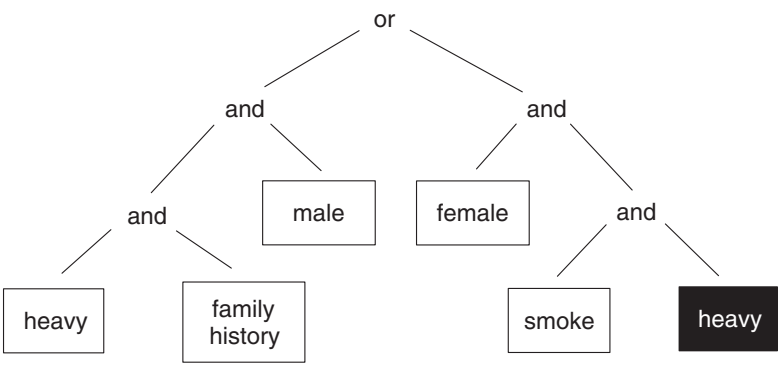

Figure 7. The tree used to generate the simulated data. Subjects are at high risk of colon cancer if they are heavy (BMI $> 25.7\ \mathrm{kg/m^2}$) males with a family history of colon cancer, or if they are female smokers (pack-years > 0) who are not heavy (BMI $\leqslant 24.2\ \mathrm{kg/m^2}$).

Table III. Operating characteristics of the tree used to generate the data (shown in Figure 7).

Age (years)	Gender	Sensitivity per cent	Specificity per cent
30–39	Female	2.0	100.0
30–39	Male	48.0	92.0
40–49	Female	53.0	87.0
40–49	Male	1.0	100.0
50–59	Female	54.0	93.0
50–59	Male	2.0	100.0
60–69	Female	50.0	95.0
60–69	Male	13.0	97.0
70–79	Female	42.0	98.0
70–79	Male	0.0	100.0

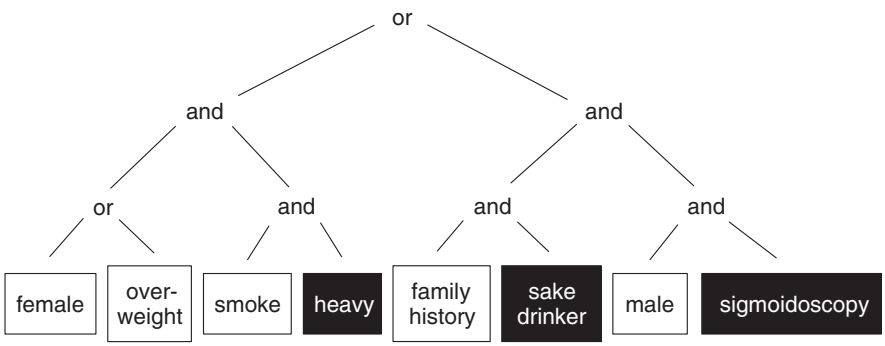

Figure 8. The logic tree fitted to the simulated data. Risk factors include smoking (pack-years >0) and not being heavy (BMI $\leqslant 24.2$ kg/m^2) for females, and a family history of colon cancer, not drinking sake (currently) and not having had a screening sigmoidoscopy (>1 year before study entry) for males.

in Figure 9; sensitivities and specificities for the logic regression model are superimposed on these plots. We see that in some strata, the stepwise logistic and logic models perform equally well, while for others, the logic regression model has significantly better discrimination. In each stratum, the fitted logic regression model performs as well or slightly better than the tree used to generate the data.

This simulation illustrates the potential value of logic regression. In settings where the high risk subpopulation is described by a complex combination of risk factors, a logic regression model yields a simple and interpretable characterization of the high risk subgroup. A logic regression model can also result in a rule that has better discrimination between cases and controls compared to the criterion that corresponds to a stepwise logistic regression model.

The operating characteristics of the tree used to generate the simulated data, shown in Table III, also have important implications. Recall that individuals falling into the subgroup described by the tree were very likely to become cases in the simulated data set (0.75 probability), while those not in this subgroup were much less likely to be cases (0.2 probability). However, the fact that a small portion of the population (15 per cent) fell into the high risk subgroup meant that a large number of cases were generated outside of the high risk subgroup.

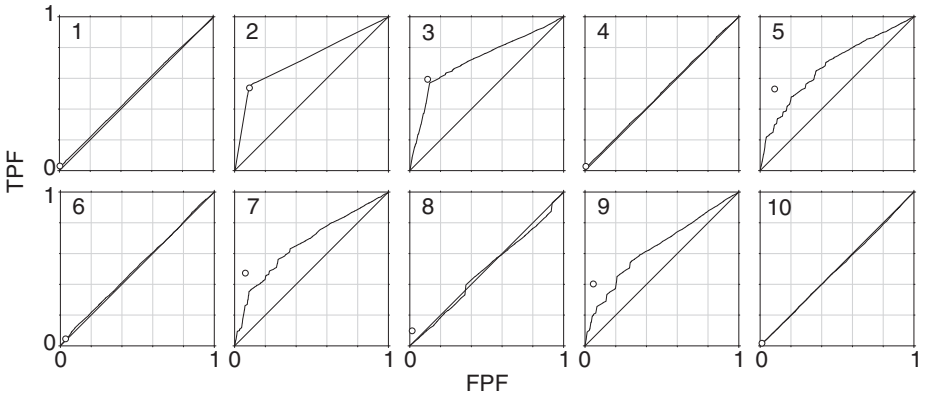

Figure 9. Operating characteristics for the stepwise logistic model fit to the simulated data. The empirical ROC curve is shown for each of the ten strata. Estimated (FPF, TPF) points for the fitted logic regression model (Figure 8) are also shown for comparison.

Thus, the stratum-specific sensitivities of the tree used to generate the data are low, but the specificities are high. This is probably not an unlikely scenario; we would expect that, if an extremely high risk subgroup existed for a particular disease, membership in the subgroup would be rare. Hence, even a small likelihood of disease outside this subgroup would mean that a rule which discriminates between cases and controls based on their subgroup membership would have low sensitivity and high specificity. As a result, any model which attempts to describe the high risk subgroup is limited by these operating characteristics.

7. DISCUSSION

Risk factors have been established for many diseases. One potential use for such information is for targeting interventions, such as screening, or for identifying groups where interventions are not needed. Risk scores based on multiple risk factors have been developed. Examples are the Framingham risk score for cardiovascular disease [17] and the Gail *et al.* breast cancer risk prediction (BCRP) model [18]. Rockhill *et al.* [19] have criticized the BCRP model because it is not very discriminatory. Many subjects who do not get disease have high risk scores while many breast cancer cases have low values prior to their disease onset. Similarly, the Framingham risk score does not discriminate well between those destined to become cases and those destined to become controls [17]. Better discriminators would clearly be more useful. We sought to identify criteria that would be discriminatory for colon cancer, with either high sensitivity or high specificity. Unfortunately, our data did not present such a criterion.

The technique that we used for extracting criteria from risk factor data is logic regression, a technique that is well suited to settings where the presence (or absence) of various combinations of risk factors yields similar risk. In our opinion, logic regression generates a much simpler characterization of the subsets of the population at high risk than does linear logistic regression, which depends on weighted averages of covariate values.

The algorithm that we implemented used the deviance ($-2 \times \log$ likelihood) as the objective function for determining the Boolean predictor variables and their co-efficients. This choice

of objective function enabled us to naturally compare logic and stepwise logistic regression. However, the deviance is not directly related to notions of accuracy associated with model-based positivity criteria (i.e. FPF, TPF, and PV). In addition, the ratio of cases to controls in the sample will affect the models selected if deviance is the objective function. It is possible that another objective function could yield better performing criteria. One possibility is to restrict attention to predictor variables that yield FPF (or TPF) values within a desirable range and to maximize TPF (or minimize FPF) within that subset. Eguchi and Copas [20] discuss such an objective function with FPF fixed at a particular value. Maximizing the area under the ROC curve associated with the fitted model has also been discussed [20, 21]. Etzioni *et al.* [9] implemented logic regression using a weighted misclassification rate, $w(1-\text{TPF})+(1-w)\text{FPF}$, as the objective function. They varied w to yield corresponding single tree models whose FPFs varied from 0 at $w=0$, to 1 at $w=1$. This approach might also be used in risk factor modelling to find Boolean criteria with desired levels of specificity (or sensitivity).

We chose thresholds or indicators corresponding to continuous covariates based on quantiles of the control distribution. Defining thresholds *a priori* according to other cut-offs may have yielded different results, although established cut-offs did not exist for the variables in our data set.

We had missing data on a number of covariates, and chose simply to drop subjects with any missing values. The amount of missing data was relatively small (6.3 per cent in controls, and 7.8 per cent in cases), especially when considered by predictor, where the maximum amount of missingness in cases occurred with multivitamin use (2.9 per cent) and in controls with non-steroidal anti-inflammatory drugs (NSAID) use (2.0 per cnt). Moreover, there was a clear lack of signal in our data. Therefore, we did not implement special procedures, such as imputation methods, to correct for bias due to missing data.

When statistical models are selected in an adaptive fashion, as is the case both for logic regression and stepwise logistic regression, selection of the 'right size' model can be quite important. In this paper we avoided this problem for logic regression by selecting the model size *a priori*. That is, we selected model sizes for logic regression that we felt would be easy to interpret. Ruczinski *et al.* [7] argue for the use of cross-validation and randomization tests to select the model that predicts best. (Software is available from: http://www.bear.fhcrc.org/~ingor/logic.) A *post hoc* cross-validation analysis we carried out suggests that, for both the one and two tree logic models for the colon cancer data, smaller models would produce at least equally good results. There is some evidence that the model sizes we chose overfit the data more than smaller models, but we felt that the amount of overfitting would not correspond to meaningful differences in the operating characteristics.

For any statistical model, selected using cross-validation or *a priori*, honestly assessing the prediction cannot be carried out on the same data that was used to fit the model. To make such an assessment, we either need a second level of cross-validation, or we need to use a separate test data set. For this analysis, we chose to split our data, using one part for training to identify predictors and estimate parameters, and the other for assessing operating characteristics of the associated criteria. This was a simple solution that worked well in our application because of the relatively large sample sizes. However, it is a somewhat inefficient use of data, and cross-validation techniques may be necessary with more limited data sets.

We have introduced logic regression, a new tree-based statistical technique for modelling binary data. Logic regression is useful for detecting subpopulations at high or low risk of disease, characterized by high-order interactions among covariates. The logic trees provide

easily interpretable descriptions of these subpopulations, and thus the methodology was well motivated for our colon cancer application. Unfortunately, our colon cancer data did not give rise to particularly high or low risk subgroups. We are confident in concluding that there is no combination of these risk factors which would be useful for targeting screening efforts in the population. However, we feel that logic regression would be useful in situations in which high-order interactions are important in determining disease risk. Our simulation demonstrates that, if there is such signal in the data, logic regression will detect it.

ACKNOWLEDGEMENTS

We would like to thank Libby Morimoto for her insights and assistance with the CCFR data, and Gary Longton for his help with the figures. This work was supported by U0ICA 074794, GM-54438, and CA 74841.

REFERENCES

1. Surveillance, Epidemiology, and End Results (SEER) Program Public-Use Data (1973–1999). National Cancer Institute, DCCPS, Surveillance Research Program, Cancer Statistics Branch, April, 2002.
2. Wooldf SH (ed.). Screening for colorectal cancer. In *Guide to Clinical Preventive Services* (2nd edn). US Preventive Services Task Force. Williams & Wilkins: Baltimore, 1996.
3. Thiis-Evensen E, Hoff GS, Sauar J *et al.* Population-based surveillance by colonoscopy: effect on the incidence of colorectal cancer. Telemark Polyp Study I. *Scandinavian Journal of Gastroenterology* 1999; **34**:414–420.
4. Newcomb PA, Storer BE, Morimoto LM *et al.* Long term efficacy of sigmoidoscopy in the reduction of colorectal cancer incidence. *Journal of the National Cancer Institute* 2003; **85**:622–625.
5. Potter JD. Colorectal cancer: molecules and populations. *Journal of the National Cancer Institute* 1999; **91**(11):916–932.
6. Colditz GA, Atwood KA, Emmons K *et al.* Harvard report on cancer prevention, Volume 4: Harvard cancer risk index. *Cancer Causes and Control* 2000; **11**:477–488.
7. Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics* 2003; **12**(3):475–511.
8. Kooperberg C, Ruczinski I, LeBlanc M *et al.* Sequence analysis using logic regression. *Genetic Epidemiology* 2001; **21**:S626–S631.
9. Etzioni R, Kooperberg C, Pepe M *et al.* Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics* 2003; **4**:523–538.
10. Hastie T, Tibshirani R, Friedman J. *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*. Springer: New York, 2001.
11. van Laarhoven PJ, Aarts EH. *Simulated Annealing*: *Theory and Applications*. Kluwer: Boston, 1987.
12. Newcomb PA, Storer BE, Morimoto LM *et al.* Long-term efficacy of sigmoidoscopy in the reduction of colon cancer incidence. *Journal of the National Cancer Institute* 2003; **95**:622–625.
13. McIntosh M, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics* 2002; **58**(3):657–664.
14. Breslow NE, Day NE. *Statistical Methods in Cancer Research*, vol. 1. International Agency for Research on Cancer: Lyon, 1980.
15. Pepe MS, Janes H, Longton GM *et al.* Limitations of the odds ratio in gauging the performance of a diagnostic or prognostic marker. *American Journal of Epidemiology* 2004; **159**:882–890.
16. Metz CE, Herman BA, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* 1998; **17**:1033–1053.
17. Cai T, Pepe MS, Lumley T *et al.* The sensitivity and specificity of markers for event times. *Working Paper No.* 188, University of Washington Working Paper Series, 2003. (Available from: http://www.bepress.com/uwbiostat/paper188.)
18. Gail MH, Brinton LA, Byar DP *et al.* Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *Journal of the National Cancer Institute* 1989; **81**:1879–1886.
19. Rockhill B, Spiegelman D, Byrne C *et al.* Validation of the Gail et al. model of breast cancer risk: prediction and implications for chemoprevention. *Journal of the National Cancer Institute* 2001; **93**(5):358–366.
20. Eguchi S, Copas J. A class of logistic-type discriminant functions. *Biometrika* 2002; **89**(1):1–22.
21. Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000; **1**(2): 123–140.

