5-23-2003

Identifying Target Populations for Screening or Not Screening Using Logic Regression

Holly Janes
University of Washington, hjanes@u.washington.edu

Margaret S. Pepe
University of Washington, mspepe@u.washington.edu

Charles Kooperberg
Fred Hutchinson Cancer Research Center, clk@fhcrc.org

Polly Newcomb
Fred Hutchinson Cancer Research Center, pnewcomb@fhcrc.org

Identifying Target Populations for Screening or Not Screening Using Logic Regression

Targeting Screening With Logic Regression

Holly Janes[1*], Margaret Pepe[1,2], Charles Kooperberg[1,2] & Polly Newcomb[3]

[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.*
[3] *Cancer Prevention Research Program, Fred Hutchinson Cancer Research Center, 1100 Fairview Avenue North, MP-900, Seattle, WA 98109, U.S.A.*

**Correspondence to: Holly Janes*
University of Washington
Department of Biostatistics
F-600 Health Sciences Building
Campus Mail Stop 357232
Seattle, WA 98195
email: hjanes@u.washington.edu
phone: (206) 322-9029
fax: (206) 543-3286

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 predicitive 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 (C-CFR), 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. 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 dataset.

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] 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 (C-CFR) 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 and End Results (SEER) registry and controls, matched to cases on age and gender, selected at random from population lists.[3] 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 (eg. ethnicity and education).[4,5]

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 fiber") along with ("male gender" or "female gender and not on post-menopausal hormones") would distinguish well between cases and controls. In Section 2 we describe logic regression, the methodology we used for combining risk factor information. This is a tree-based statistical technique for identifying subsets of the population defined by such Boolean functions of binary coded risk factors and is therefore well suited to our purposes. In Section 3 we describe the C-CFR 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 risk factor combinations do discriminate cases from controls. In Section 6 we demonstrate this with a simulated dataset. 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

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

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. Ruczinski, Kooperberg and LeBlanc[6] provide a detailed description of logic regression and the simulated annealing algorithm used to fit it. The output 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.

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 Kooperberg et al.[7] for a successful application of logic regression in this setting. We suspect that disease risk factors may behave similarly. Etzioni et al.[8] 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.[6] provide further examples of applications of logic regression.

In addition to the specification of the scoring function, the 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 (3 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.

For a given model size, the selection of the best logic trees L_j is a nontrivial optimization problem. The logic regression algorithm that we implemented employs a simulated annealing algorithm. Simulated annealing[9] is a stochastic optimization algorithm similar to the Metropolis-Hastings algorithm for Markov chain Monte Carlo.[6] 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 to 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.[10] Response rates were high (80% for cases, 71% for controls).[3]

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. For BMI and height, different thresholds were chosen for men and women. Sigmoidoscopy screening history was defined as screening more than one year prior to enrollment in the study. 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% 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 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) and specificity (1 - false positive fraction) 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, $P(D = 1 \mid L_1, \ldots L_P)$, or equivalently the linear predictor, exceeding a threshold c:

$$\text{“positive”} = \text{“}\beta_1 L_1 + \ldots + \beta_P L_P > c\text{”}.$$

Such decision criteria are known to be optimal.[11] The associated true and false positive fractions,

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

and

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

are quantities derived from cases and controls, respectively. A plot of $(FPF(c), TPF(c))$ displays the range of operating characteristics attainable with the risk factors. This plot is known as the Receiver Operating Characteristic (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, subjects who did not meet the criterion could forego screening without losing many cases to 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. We used the SEER incidence rates for colorectal cancer (denoted by ρ) and the following relationships to calculate predictive values (PV):

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

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 $\tau = \text{Prob}(\mathit{positive})$, in order to assess the impact of using such a criterion in the population. We calculate τ with the formula:

$$\tau = \rho TPF + (1-\rho)FPF.$$

4.3 Stratum Specific Performance

As is typical of many case-control studies, the C-CFR 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.[12] We categorized age into 5 categories, which along with gender defines 10 strata. A stratum-specific intercept, α_s for $s = 1, \ldots 10$ was included in the model

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

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.

The assessment of criteria such as "$\beta_1 L_1 + \ldots + \beta_P L_P > c$" is straightforward within strata. We can calculate the $(FPF(c), TPF(c))$ values using the cases and controls within each stratum. Predictive values can be calculated with stratum-specific incidence rates (available from SEER). It is not clear how best to summarize these operating characteristics across strata, particularly if they vary amongst strata. Therefore we report the 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% confidence interval associated with the tree are $\exp(\hat{\beta}_1) = 2.9$ and (2.1, 3.9), respectively, with p-value $< .001$.

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.[4] 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% confidence interval = (2.0, 4.5), p-value $< .001$). 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 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 1). Again, we note that performance is similar with the validation and training datasets, although there is more statistical variability with the smaller validation set, as expected. The sensitivities, averaging about 45-50%, are not very high. We certainly could not use this criterion to 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% across the strata. However, it may not be appropriate to use this criterion for targeting intense screening encouragement efforts either: about 24% 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 1, 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:

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

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, $(1 - TPF)$ 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.

We now turn to the population performance of the criterion. Table 1 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% across the strata. Note that the incidence of colon cancer is very low, ranging from about 20 per 100,000 per year in 40-50 year old women to 364 per 100,000 per 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 1). The highest value is seen in 70-79 year old females where the incidence of colon cancer is estimated to be 8.1 per 1,000 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. It is possible that smaller or larger models would perform slightly better. However, we do not expect them to have substantially different 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 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 most easily interpretable model.

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's 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% across strata, with corresponding average sensitivities of 25%. If these numbers are accurate, it appears that 25% of cases could be identified for screening with the criterion without referring more than 11% of non-diseased subjects for unnecessary screening. The most sensitive criterion averaged 83% with specificities that average 33% across the strata. If these numbers are accurate, we could save 33% of controls from unnecessary screening while continuing to screen the majority of cases.

As more trees are added to the model, this creates a broader range of criteria that can be investigated. There are, in fact, P criteria that are formed from the linear predictor $\beta_1 L_1 + \ldots + \beta_P L_P$. Assume without loss of generality that $\beta_1 > \beta_2 > \ldots > \beta_P$. The constants $c_1 = \beta_1 + \ldots + \beta_P, \ldots,\ c_{P-1} = \beta_{P-1} + \beta_P,\ c_P = \beta_P$ define the P criteria that are nested in the sense that all subjects positive by criterion p are also positive by lower order criteria $q < p$. In general, the associated operating characteristics are represented as P points along an ROC curve.

5.3 Comparison with Linear Logistic Regression

We fit a linear logistic model to the C-CFR data. A stepwise algorithm yielded the results shown in Table 2. Covariates whose statistical significance was $p < .2$ 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

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

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

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

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.[13] The average curve

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

where $\overline{a} = \sum_{s=1}^{S} a_s/S$ and $\overline{b} = \sum_{s=1}^{S} b_s/S$, 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. The linear logistic model criteria are more complex than those from logic regression. One needs to calculate the linear score (3), i.e. a weighted average of risk factors, and compare it with the chosen threshold. The logic trees, on the other hand, produce more easily defined subsets of the population that do not involve weighted averaging, although this comes at a cost of some constraints on risk factor parametrization. In this dataset, 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 Dataset

In order to validate the use of logic regression in a setting in which combinations of covariates are important for predicting disease, we simulated such a dataset. 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 ≤ 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 3. 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% to 2%). The specificities are high, a result of the rarity of the high risk subgroup (87% to 100%).

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 dataset ($N = 78,000$). The stratum-specific empirical ROC curves for the logistic model are shown 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 3, also have important implications. Recall that individuals falling into the subgroup described by the tree were very likely to become cases in the simulated dataset (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%) fell into the high risk subgroup meant that a large number of cases were generated outside of the high risk subgroup. 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[14] and the Gail et al. breast cancer risk prediction (BCRP) model.[15] Rockhill et al.[16] 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.[14] 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, the criteria that are generated from logic regression are more intuitively appealing than those from linear logistic regression that depend 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 predictive values). 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[17] 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.[17,18] Etzioni et al.[8] implemented logic regression using a weighted misclassification rate, $w(1 - TPF) + (1 - w)FPF$, 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 modeling 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 meaningful cutoffs may have yielded different results. Another option would be to include several distinct sets of threshold indicators and let the algorithm choose the most discriminating ones.

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 were easy to interpret. Ruczinski et al.[6] 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.) Some limited cross-validation that we carried out suggests that, for both the one and two tree logic models for the colon cancer data and for the simulated data model, slightly smaller models would produce at least equally good results.

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 dataset. 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 is simple, though a somewhat inefficient use of data.

Acknowledgments: We would like to thank Libby Morimoto for her insights and assistance with the C-CFR data. 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, released April 2002, based on the November 2001 submission.

[2] U. S. Preventive Services Task Force. Screening for colorectal cancer. In S. H. Woolf (ed.), *Guide to Clinical Preventive Services*, 2nd ed. Williams and Wilkins: Baltimore, 1996.

[3] Newcomb PA, Storer BE, Morimoto LM, et al. Long term efficacy of sigmoidoscopy in the reduction of colorectal cancer incidence. *J Natl Cancer Inst* 2003;**85**:622-5.

[4] Potter JD. Colorectal cancer: molecules and populations. *J Natl Cancer Inst* 1999;**91**(11):916-32.

[5] 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**:477-88.

[6] Ruczinski I, Kooperberg C, LeBlanc M. Logic regression. *Journal of Computational and Graphical Statistics*, in press.

[7] Kooperberg C, Ruczinski I, LeBlanc M, et al. Sequence analysis using logic regression. *Genetic Epidemiology* 2001;**21**:S626-31.

[8] Etzioni R, Kooperberg C, Pepe M, et al. Combining biomarkers to detect disease with application to prostate cancer. *Biostatistics*, in press.

[9] van Laarhoven PJ, Aarts EH. *Simulated Annealing: Theory and Applications.* Kluwer: Boston, 1987.

[10] Newcomb PA, Haile R, Anton-culver H, et al. The colorectal cancer family registry. *Cancer Epidemiology Biomarkers and Prevention* 2002;**11**(10)2:1222s.

[11] McIntosh M, Pepe MS. Combining several screening tests: optimality of the risk score. *Biometrics* 2002;**58**(3):657-64.

[12] Breslow NE, Day NE. *Statistical Methods in Cancer Research*, Volume 1. International Agency for Research on Cancer: Lyon, 1980.

[13] Metz CE, Herman BA, Shen J. Maximum likelihood estimation of receiver operating characteristic (ROC) curves from continuously distributed data. *Statistics in Medicine* 1998;**17**:1033-53.

[14] Cai T, Pepe MS, Lumley T, et al. The sensitivity and specificity of markers for event times. *UW Working Paper Series* 2003; Working Paper 188. Available from: http://www.bepress.com/uwbiostat/paper188.

[15] Gail MH, Brinton LA, Byar DP, et al. Projecting individualized probabilities of developing breast cancer for white females who are being examined annually. *J Natl Cancer Inst* 1989;**81**:1879-86.

[16] Rockhill B, Spiegelman D, Byrne C, et al. Validation of the Gail et al. model of breast cancer risk: prediction and implications for chemoprevention. *J Natl Cancer Inst* 2001;**93**(5):358-66.

[17] Eguchi S, Copas J. A class of logistic-type discriminant functions. *Biometrika* 2002;**89**(1):1-22.

[18] Pepe MS, Thompson ML. Combining diagnostic test results to increase accuracy. *Biostatistics* 2000;**1**(2):123-40.

Figures

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.

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 kg/m^2 for females, > 27.3 kg/m^2 for males); p.m. hormones (women post-menopause ever taking hormones for more than 6 months).

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, 16.0.

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 (≤ 2 servings per week); screening sigmoidoscopy (> 1 year before study entry); NSAID use (> 0.25 months using non-steroidal anti-inflammatory drugs); college education (some college education).

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 1b. Values for most sensitive ($\circ$) and most specific ($\triangle$) criteria are displayed.

Figure 6 Operating characteristics associated with the linear logistic model. The average ROC curve is shown, $ROC(t) = \Phi(\overline{a} + \overline{b}\Phi^{-1}(t))$ with $\overline{a} = .55$ and $\overline{b} = 1.05$. Estimated (FPF, TPF) points for the single tree logic regression model are also 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 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 ≤ 24.2 kg/m^2).

Figure 8 The logic tree fitted to the simulated data. Risk factors include smoking (pack-years > 0) and not being heavy (BMI ≤ 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.

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.

Tables

Table 1 Operating characteristics for the single tree model (Figure 2) for the colon cancer data. (A) using the training dataset; (B) using the validation dataset.

Table 2 Results of a linear logistic regression model fit to the colon cancer data. Age and gender were included in the model.

Table 3 Operating characteristics of the tree used to generate the data (shown in Figure 7).

Table 1

(A) Training Data

Age (years)	Gender	Number of Cases	Sensitivity %	Number of Controls	Specificity %	Odds Ratio	Criterion Positive %	1 Year Positive PV %
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

Table 1

(B) Validation Data

Age (years)	Gender	Number of Cases	Sensitivity %	Number of Controls	Specificity %	Odds Ratio	Criterion Positive %	1 Year Positive PV %
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

Table 2

	Odds Ratio	95% 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		
per 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)

Table 3

Age (years)	Gender	Sensitivity %	Specificity %
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
79-79	Male	0.0	100.0

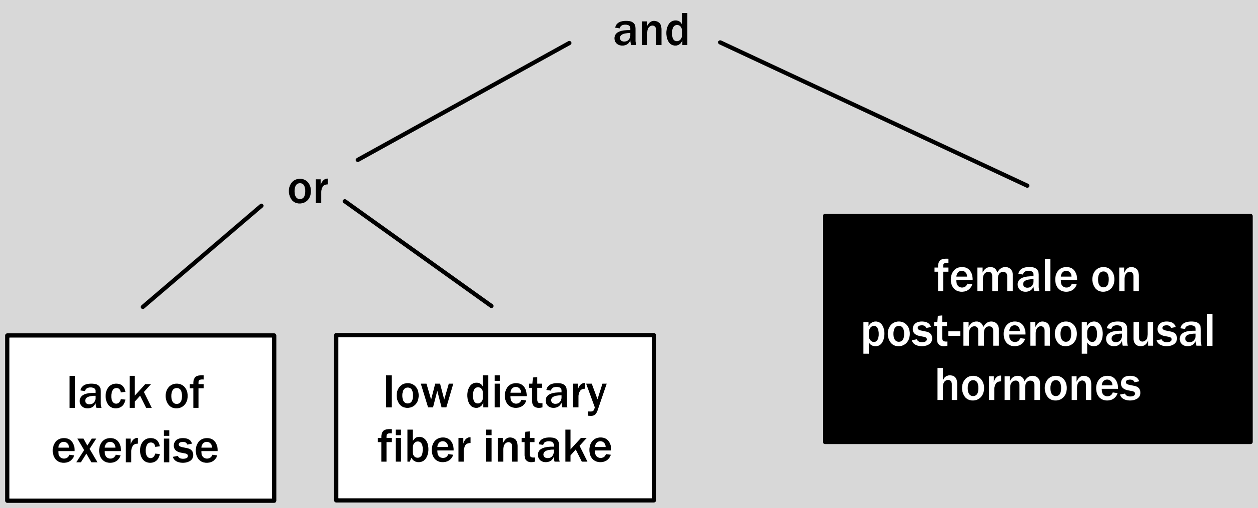


Figure 1 Example of a logic tree that evaluates to 1 if the Boolean expression illustrated is true. White letters on black denote negation of the entry.

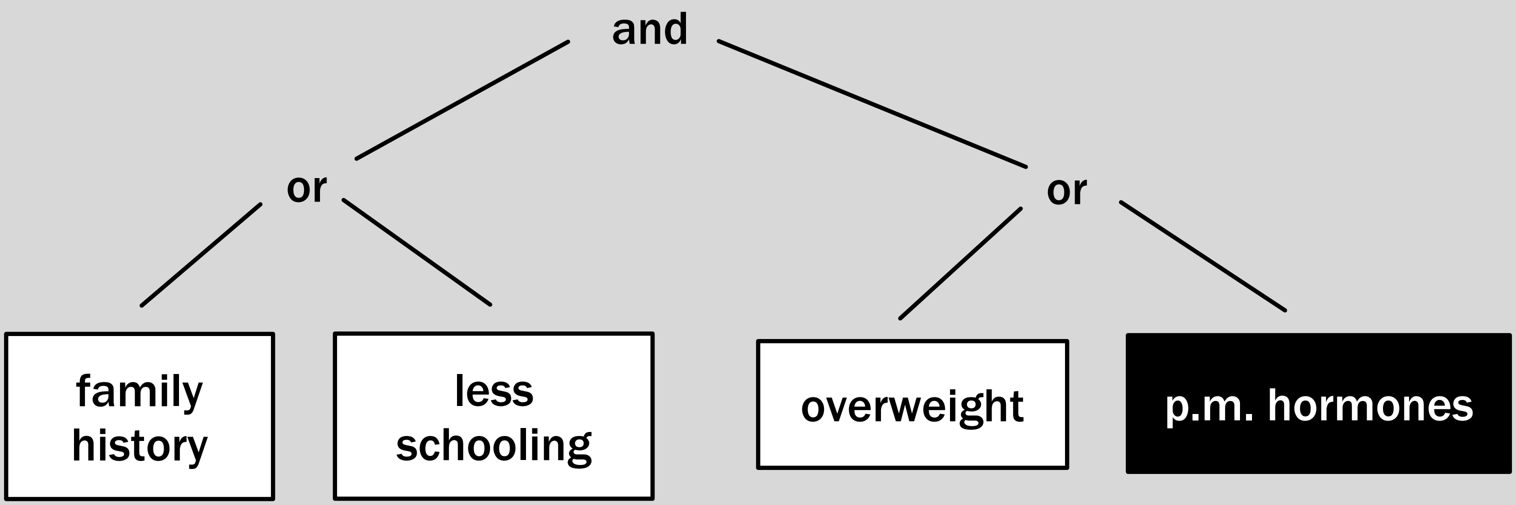


Figure 2 The single tree, L_1, fitted to the colon cancer data. The risk factors included are: family history (yes/no); less schooling (high school education or less); overweight (body mass index > 26.6 kg/m^2 for females, > 27.3 kg/m^2 for males); p.m. hormones (women post-menopause ever taking hormones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

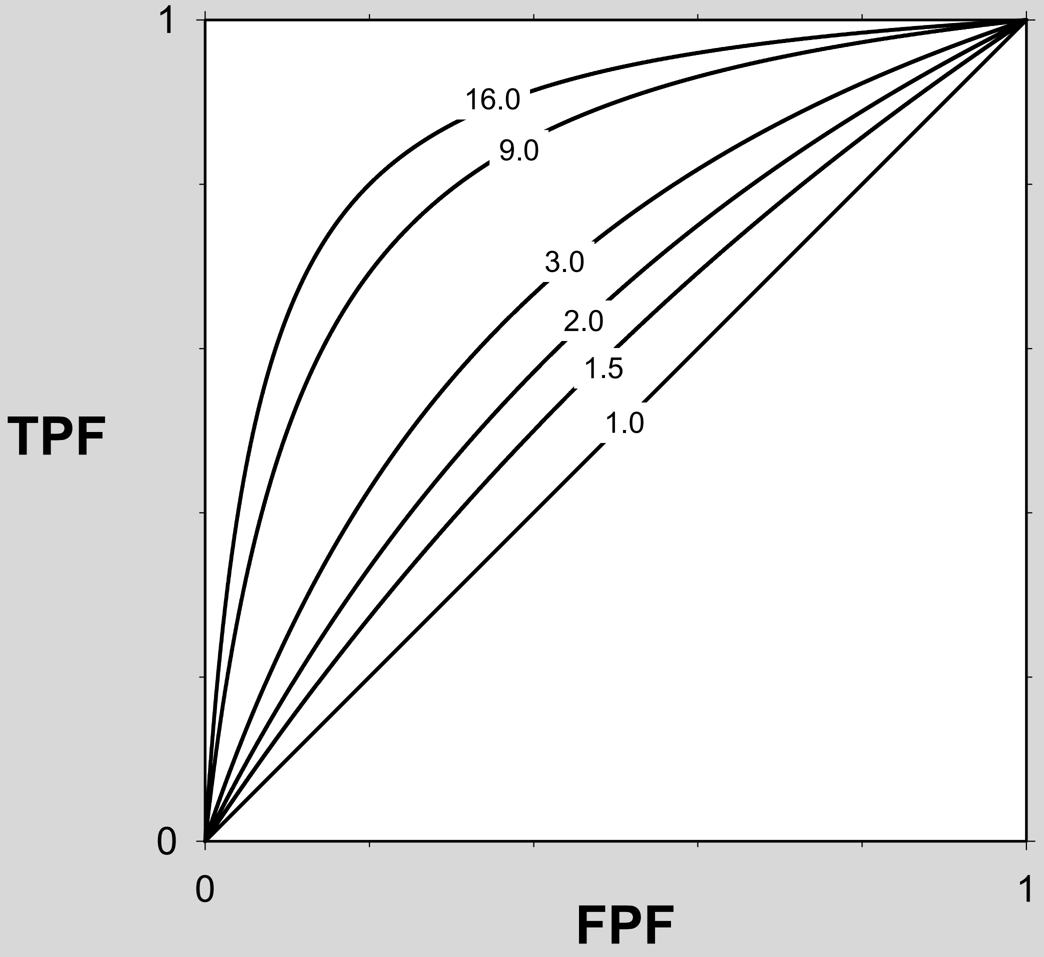
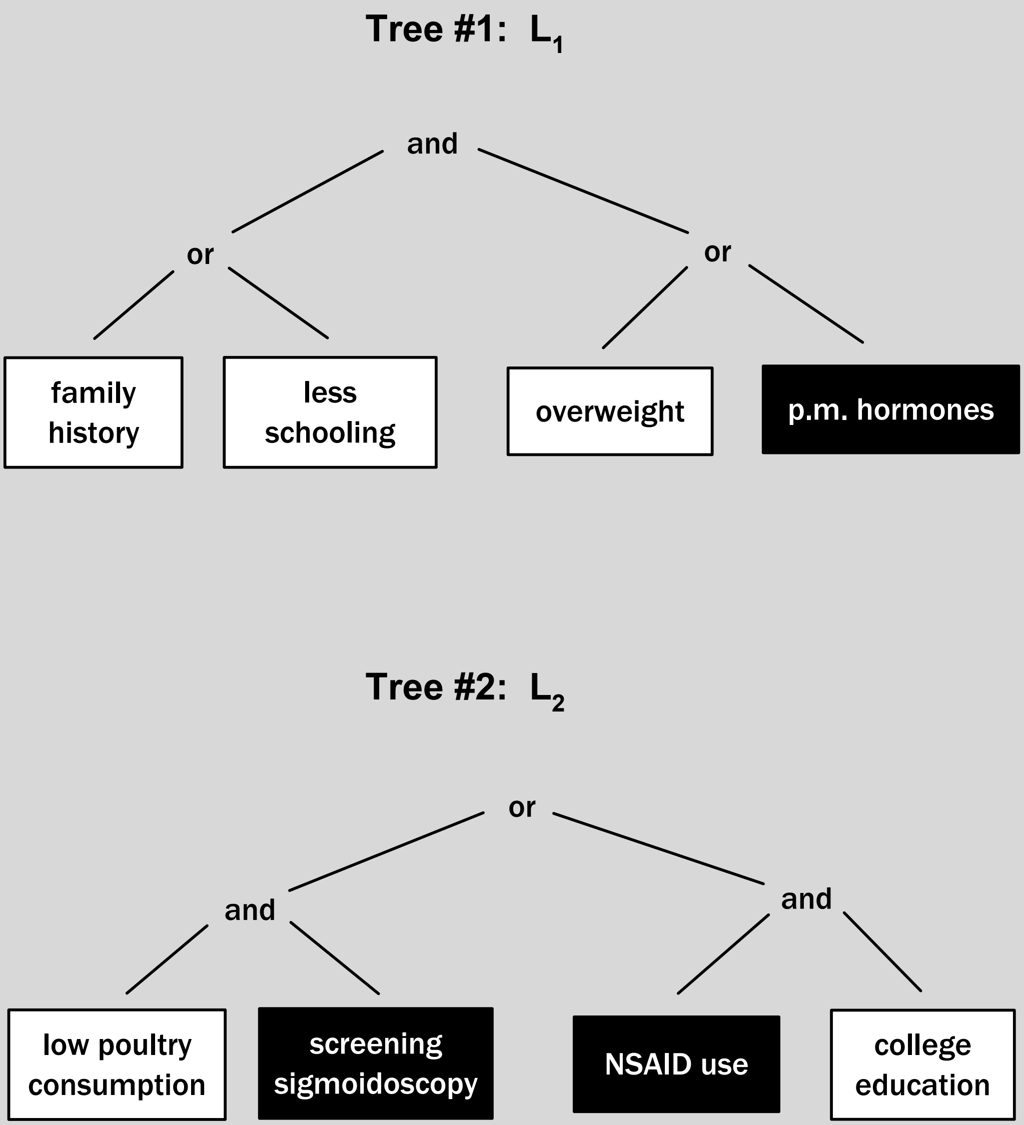


Figure 3 Contour plots for the odds ratio. (FPF, TPF) combinations that yield equal values for the odds ratio are connected. Shown are contours for odds ratios of 1.0, 1.5, 2.0, 3.0, 9.0, 16.0.



Tree #1: L_1

Tree #2: L_2

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 $\beta_1 L_1 + \beta_2 L_2 = 1.096L_1 + 0.777L_2$. Variables in L_1 are described in Figure 1. Variables in L_2 are: low poultry consumption ($\leq$ 2 servings per week); screening sigmoidoscopy (> 1 year before study entry); NSAID use (> 0.25 months using non-steroidal anti-inflammatory drugs); college education (some college education).

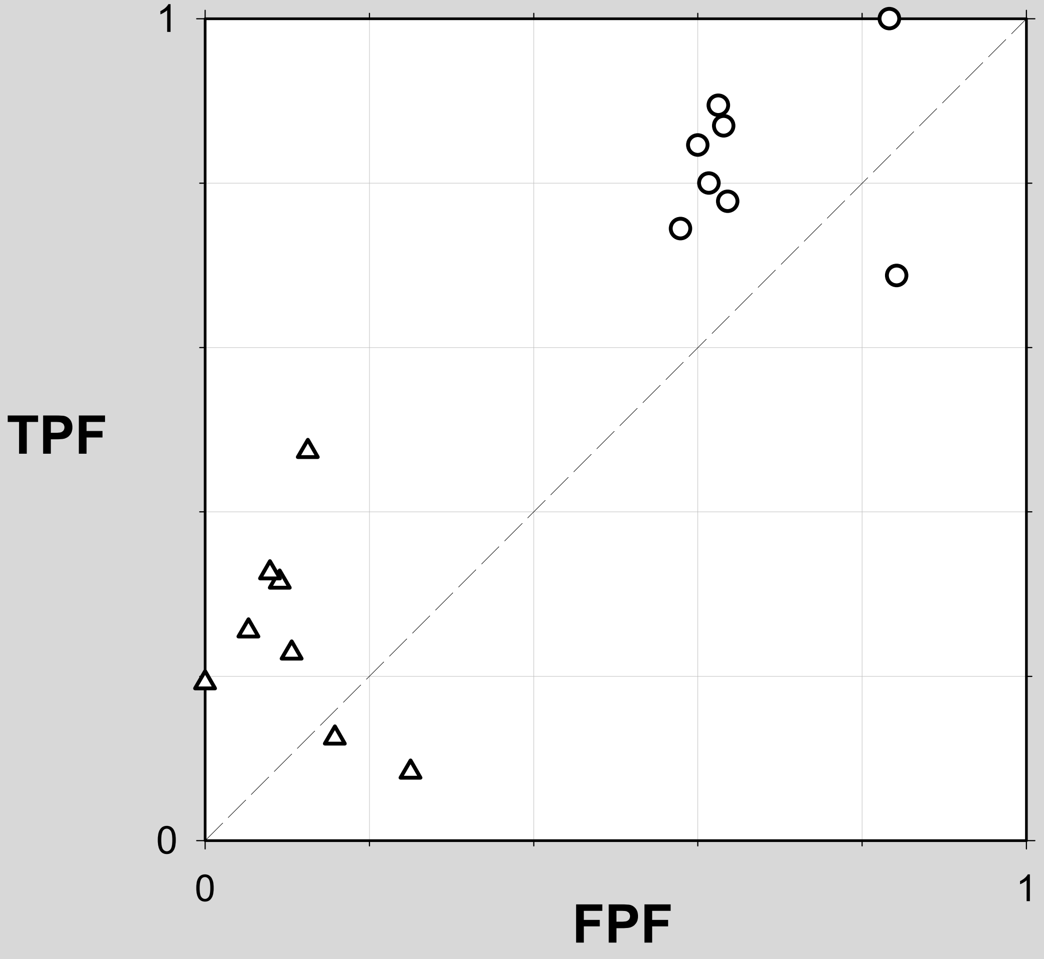


Figure 5 Operating characteristics for criteria based on the two-tree model. Each point represents a stratum with numbers of cases and controls shown in Table 1b. Values for most sensitive (○) and most specific (△) criteria are displayed.

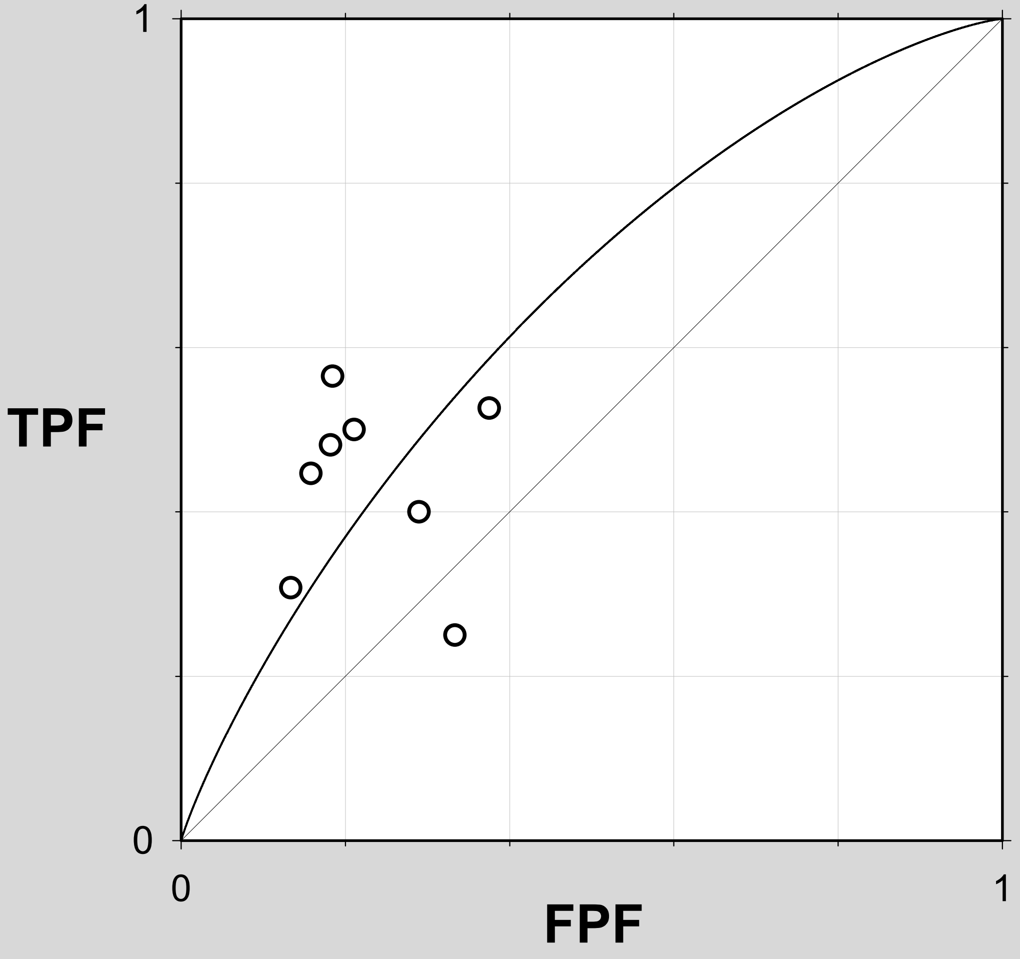


Figure 6 Operating characteristics associated with the linear logistic model. The average ROC curve is shown, ROC(t) = $\Phi(\bar{a} + \bar{b}\Phi^{-1}(t))$ with $\bar{a}$ = .55 and $\bar{b}$ = 1.05. Estimated (FPF, TPF) points for the single tree logic regression model are also shown.

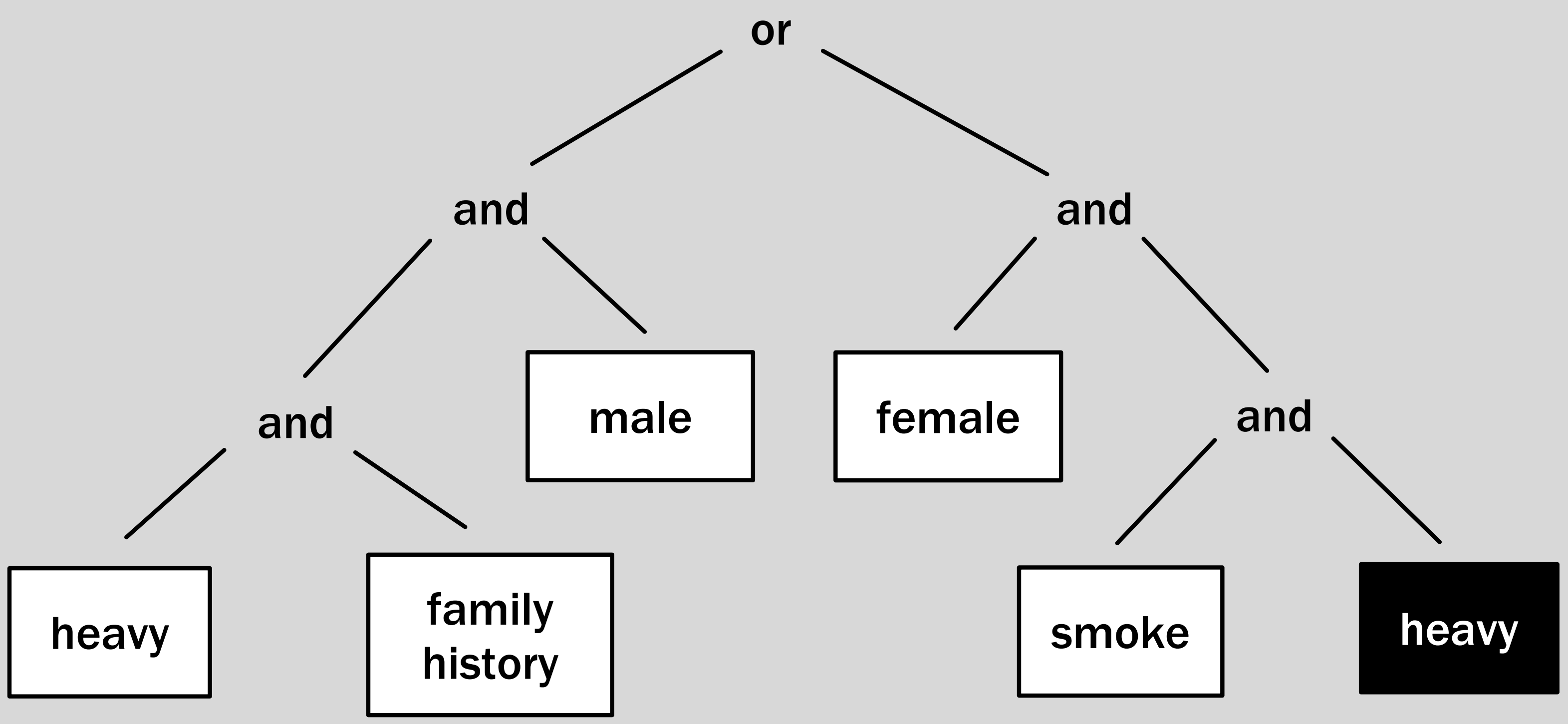


Figure 7 The tree used to generate the simulated data. Subjects are at high risk of colon cancer if they are heavy (BMI > 25.7 kg/m^2) males with a family history of colon cancer, or if they are female smokers (pack-years > 0) who are not heavy (BMI $\leq$ 24.2kg/m^2).

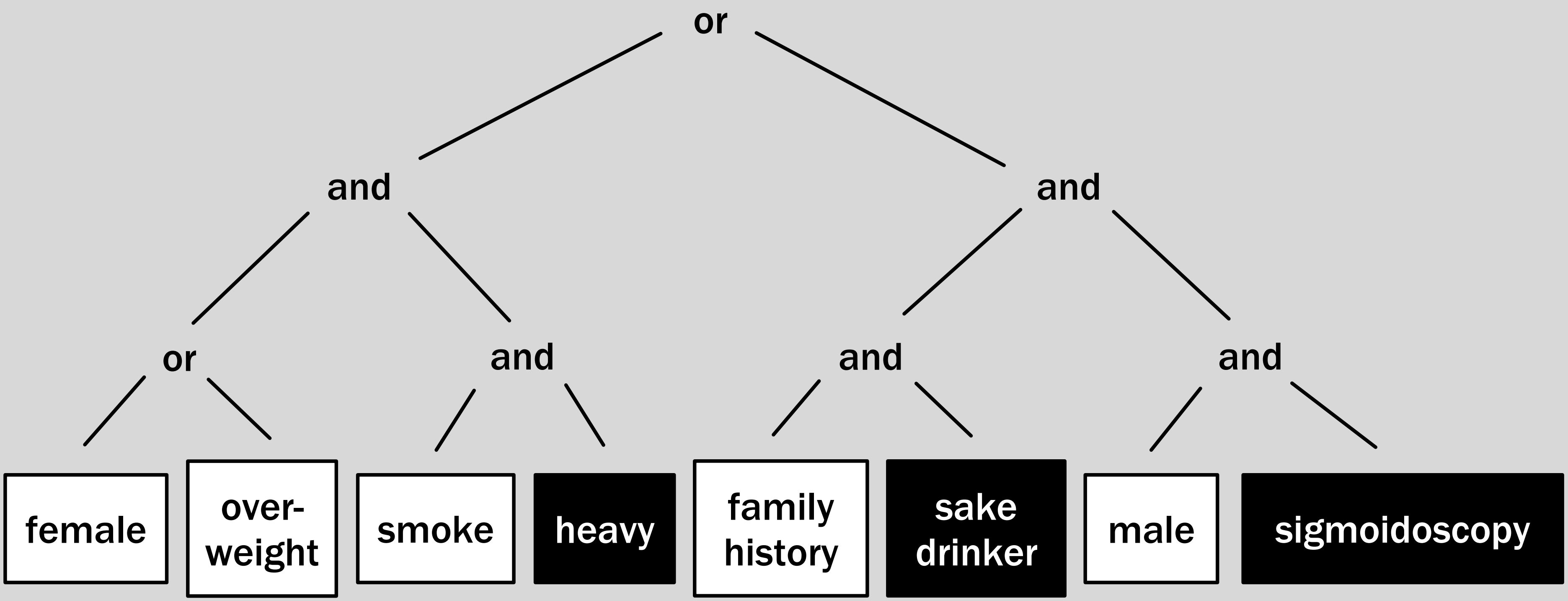


Figure 8 The logic tree fitted to the simulated data. Risk factors include smoking (pack-years > 0) and not being heavy (BMI ≤ 24.2 kg/m^2) for females, and a family history of colon cancer, not drinking sake (currently) and not having had a screening sigmoidoscopy (> 1 year before study entry) for males.

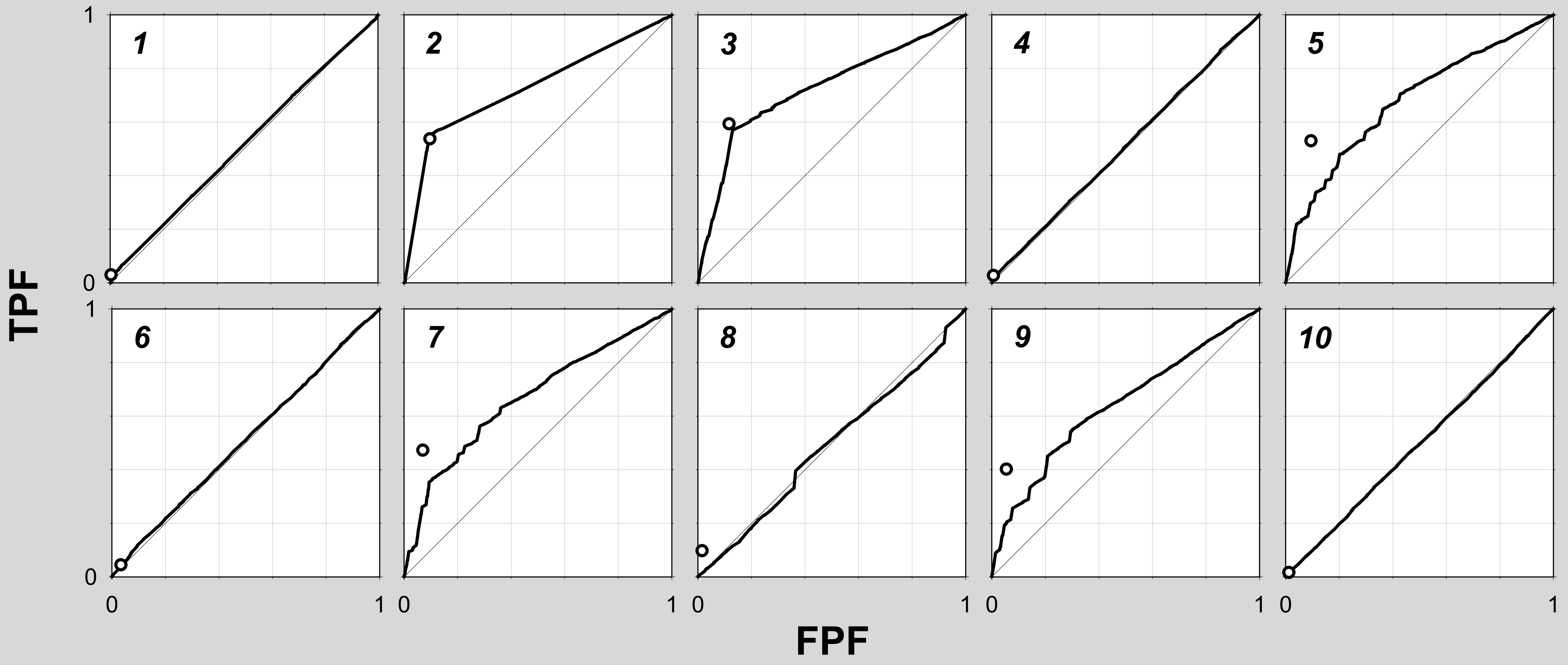


Figure 9 Operating characteristics for the stepwise logistic model fitted to the simulated data. The empirical ROC curve is shown for each of the ten age and gender strata. Estimated (FPF, TPF) points for the fitted logic regression model (Figure 8) are also shown for comparison.