Biostatistics (2011), **12**, 1, *pp.* 68–86
doi:10.1093/biostatistics/kxq047
Advance Access publication on July 27, 2010

Adaptive index models for marker-based risk stratification

LU TIAN*

Department of Health Research & Policy, Stanford University, Stanford, CA 94305, USA

lutian@stanford.edu

ROBERT TIBSHIRANI

Departments of Health Research & Policy and Statistics, Stanford University, Stanford, CA 94305, USA

tibs@stat.stanford.edu

SUMMARY

We use the term "index predictor" to denote a score that consists of K binary rules such as "age $>$ 60" or "blood pressure $>$ 120 mm Hg." The index predictor is the sum of these binary scores, yielding a value from 0 to K. Such indices as often used in clinical studies to stratify population risk: They are usually derived from subject area considerations. In this paper, we propose a fast data-driven procedure for automatically constructing such indices for linear, logistic, and Cox regression models. We also extend the procedure to create indices for detecting treatment–marker interactions. The methods are illustrated on a study with protein biomarkers as well as a large microarray gene expression study.

Keywords: Degree of freedom; Index predictor; International prognostic index.

1. INTRODUCTION

When predicting a phenotype such as clinical response or survival time from a set of biomarkers, an "index predictor" is sometimes used. It consists of a set of binary rules such as "marker $X_k \geqslant c_k$" or "marker $X_k < c_k$" for each of K markers. For each observation, we add up the binary scores yielding an index s taking values in $\{0, 1, \ldots, K\}$. This has the advantage of simplicity: It is easy to state and interpret and also can capture situations where prognostic effects are shared by multiple markers. A popular example is the international prognostic index (IPI) used for risk classification in non-Hodgkin's lymphoma (TIN-HsLPF, 1993). The IPI consists of one point for each of the following:

- age greater than 60 years,
- stage III or IV disease,
- elevated serum lactate dehydrogenase (LDH) (>1),

*To whom correspondence should be addressed.

- Eastern Cooperative Oncology Group (ECOG)/Zubrod performance status of 2, 3, or 4, and
- more than 1 extranodal site.

The resulting score lies in 0–5, with higher scores indicating greater risk. Sometimes the IPI is further simplified into 2 or 3 categories as (low, high) or (low, medium, and high) for risk stratification.

An example is shown in Figure 1. Shown are the survival curves from a set of patients with non-Hodgkin's lymphoma for each of the levels of the IPI. There is a clear separation in the groups.

In this paper, we propose a method for adaptively constructing an index predictor from training data. We also return to this example and demonstrate that our proposal can reconstruct the IPI empirically.

This paper is organized as follows. In Section 2, we introduce the adaptive index model (AIM) and our algorithm for its estimation. We discuss an example in which protein biomarkers are used to predict the presence of ovarian cancer. In Section 3, we discuss the AIM for survival outcomes, using Cox proportional hazards model, and illustrate how the AIM procedure can rediscover the IPI discussed above. We extend the AIM procedure to look for interactions between markers and a binary treatment factor in Section 5. We also discuss the construction of "surrogate markers." The degrees of freedom of the AIM procedure are discussed in Section 6. There are clear connections to other methods such as CART (Breiman *and others*, 1984), PRIM (Friedman and Fisher, 1999), their more recent refinements in LeBlanc *and others* (2002, 2005), boosted trees (see, e.g. Friedman *and others*, 2000; Friedman, 2001), and the logic regression (Ruczinski *and others*, 2003). We discuss these and make some concluding remarks in Section 7.

2. Adaptive index models

Consider a supervised learning problem (regression or generalized regression) with data $(\mathbf{x}_i, y_i)$, $i = 1, 2, \ldots N$. Here, $\mathbf{x}_i$ is a p-vector of predictor variables and y_i in an outcome variable. The 3 major applications that we consider are the linear regression, logistic regression for binary data where $y_i \in \{0, 1\}$, and Cox regression for survival outcome $y_i = (T_i, \delta_i)$, where T_i is a right censored survival time and δ_i is the censoring indicator. Denote the log likelihood or log partial likelihood by $\ell(\eta; \mathbf{X}, \mathbf{y})$, where η is the usual linear combination of predictors. For example, η is the linear predictor in simple linear regression,

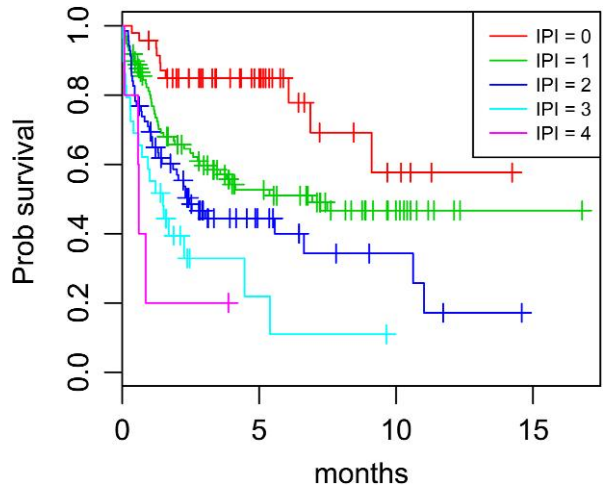

Fig. 1. Survival curves from a set of patients with non-Hodgkin's lymphoma for each of the levels of the IPI.

the log odds in logistic regression, and the log hazard in proportional hazards regression. We consider an *index model* in the form of

$$\eta = \beta_0 + \beta \times \sum_{k=1}^{K} I(\widetilde{X}_k^* \leqslant c_k), \tag{2.1}$$

with $K \leqslant p$. The predictors $\widetilde{X}_k^*$s are from the set $\mathcal{S} = \{\pm X_1, \pm X_2, \ldots, \pm X_p)$ and the corresponding cutpoints c_k are chosen in a forward stepwise manner to approximately maximize $\ell(\eta; \mathbf{X}, \mathbf{y})$, where $X_1, \ldots, X_p$, are the original p predictors of interest. The result is a simple "index" predictor $s = \sum_{k=1}^{K} I(\widetilde{X}_k^* \leqslant c_k)$, which is just a count ranging from 0 to K. By allowing $\widetilde{X}_k^*$ to equal $-X_j$, we effectively allow cuts of the complementary form $X_j \geqslant c_j$.

What makes our procedure attractive is the fact that as we change the cutpoint c_k, updating formulas can be derived for the score test for testing $\beta = 0$ in model (1). Next we give detailed updating scheme for linear and logistic models.

Our model is

$$\eta_i = \beta_0 + \beta \sum_{k=1}^{K} I(\widetilde{x}_{ki}^* \leqslant c_k),$$

where $\widetilde{x}_{ki}^*$ is the ith observation of $\widetilde{X}_k^*$. $\eta_i = E(Y_i|\mathbf{x}_i)$ in the linear model, while $\eta_i = \log\{p_i/(1-p_i)\}$ in the logistic model, where $p_i = \text{Prob}(y = 1|\mathbf{x}_i)$. Suppose that we have a score $s = \sum_{k=1}^{J} I(\widetilde{X}_k^* \leqslant c_k)$ and want to decide whether to add a binary score $z = I(X^* \leqslant c)$, where $X^* \in \mathcal{S}$. Hence, we fit a regression model with $\eta = \beta_0 + \beta(s + z)$ and test $\beta = 0$. Letting $w = s + z$ and $\hat{\mu}_0 = \bar{y}$, the average of $\{y_i, i = 1, \ldots, N\}$, we have the score vector and information matrices for testing $\beta = 0$:

$$(U_1, U_2) = \left(\sum_{i=1}^{N} (y_i - \hat{\mu}_0), \sum_{i=1}^{N} w_i (y_i - \hat{\mu}_0) \right)$$

$$\text{and} \quad I = \begin{pmatrix} I_{11} & I_{12} \\ I_{21} & I_{22} \end{pmatrix} = \begin{pmatrix} n\hat{v}_0 & \hat{v}_0 \sum_{i=1}^{N} w_i \\ \hat{v}_0 \sum_{i=1}^{N} w_i & \hat{v}_0 \sum_{i=1}^{N} w_i^2 \end{pmatrix},$$

where $\hat{v}_0$ is the empirical variance of $\{y_i\}$ in both the linear and the logistic models. The score test statistics is $U_2/V^{1/2}$, where $V = I_{22} - I_{12}^2/I_{11}$.

Without loss of generality, we assume that the observations are sorted according to the candidate marker $X^* = \{x_i^*, i = 1, \ldots, N\}$, that is $-\infty = x_0^* < x_1^* < x_2^* < \cdots < x_N^*$, throughout the paper. We have the following updating formulas as the cutpoint c moves from x_{i-1}^* to x_i^*:

$$U_2 \leftarrow U_2 + (y_i - \hat{\mu}_0),$$

$$I_{12} \leftarrow I_{12} + \hat{v}_0 \quad \text{and} \quad I_{22} \leftarrow I_{22} + \hat{v}_0(1 + 2s_i).$$

Thus, we can scan through all possible cutpoints for a given predictor in just $O(n)$ operations. We summarize the algorithm below, called "AIM."

AIM procedure

1. Begin with $k = 0$, $s = 0$.
2. For $k = 1, 2, \ldots, K$, update $s \leftarrow s + I(X_j \geqslant c_j)$ or $s \leftarrow s + I(X_j \leqslant c_j)$, where (j, c_j) maximize the score test statistics over the markers not yet entered.

In practice, we set the maximum model size K to, say, 10 or 20 and estimate the best model size by cross-validation.

Figure 2 shows the run time in seconds of the AIM procedure for logistic regression with various combinations of n, p and the maximum model size K. In the left and middle panels, we run the algorithm until $K = 20$ terms have been added. In the right panel, we have fixed p at 100. We see that the algorithm is remarkably fast and scales roughly linearly in n, p, and K.

Similar to other "stepwise" procedure, the proposed AIM algorithm does not necessarily converge to the global optimal solution, which maximizes the score test statistics. However, identifying the global optimal solution is non-deterministic polynomial-time hard and numerically infeasible when the number of features is greater than 3 or 4. On the other hand, based on our limited experience and simulation studies reported later, the AIM algorithm often locates the optimal or near optimal solution in practice. In post-AIM analysis, one may iteratively readjust the split point for each of the binary rule, while keeping others fixed, via a fast "backfitting procedure." This additional step ensures that the final solution is "locally" optimal and increases the chance of finding the global maximizer. In some cases, it can also be helpful to preprocess the data with method such as supervised principal components analysis (see Paul *and others*, 2008). We also note that this model is related to "boosted trees" (see, e.g. Friedman *and others*, 2000; Friedman, 2001) in the case where the trees are stumps (single split trees). In AIM, we further constrain all the stumps to share the same multiplier. In the case of a single predictor, the AIM procedure is equivalent to

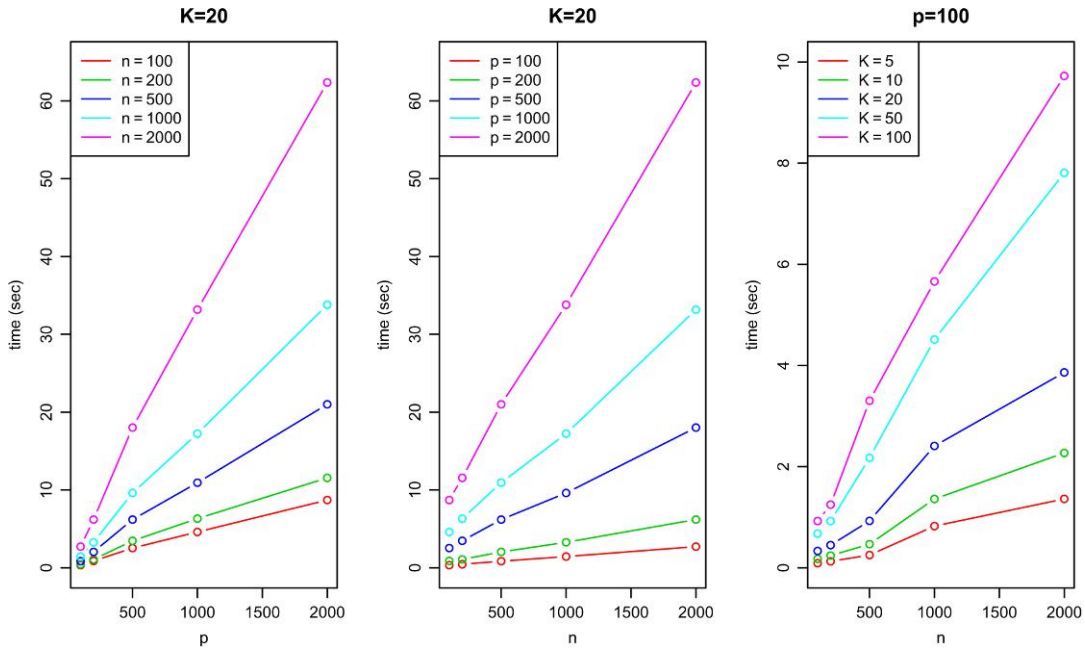

Fig. 2. Timings (seconds) for the AIM procedure for various values of n, p, and the number of terms K.

simply splitting the sample into 2 groups with a threshold value, which maximizes a 2-group comparison test statistics, and thus is a single split tree.

2.1 *Ovarian cancer data*

The data for this example are taken from Fredriksson *and others* (2008). It consists of 20 blood protein biomarkers in each of 2 groups: healthy and ovarian cancer patients. There are 20 patients in each group. We applied the AIM procedure with a maximum of 10 biomarkers. Ten-fold cross-validation was used to assess the model prediction, producing the curves in Figure 3. In the figure, we have used 2 different methods for making predictions. The "logit" method fits a logistic regression in the training set to the estimated index s and then makes predictions in the test set using the estimated probabilities from the fitted model. The "thresholding method" finds the value c that produces the fewest errors in predicting y (or $1-y$) as $I(s \leqslant c)$ in the training set and then uses this threshold level c to make predictions in the test set. Both methods yield error rates of about 10–15% with 2–3 markers and perform better than nearest shrunken centroids procedure that was used in the original paper. Also shown are the error rates for standard forward stepwise logistic regression. The AIM score s has the form of $I(X_7 \leqslant 11.8) + I(X_{16} > 9.0) + I(X_9 \leqslant 12.0)$. The optimal cutpoint is $s \leqslant 2$, so the prediction rule classifies to the cancer group if all 3 biomarkers fall in their "red" regions. Figure 4 shows that these 3 markers over the training set, with red points indicating that the corresponding condition (such as "$X_7 \leqslant 11.8$") is satisfied. Figure 5 shows a schematic of the final model.

Figure 6 shows the result of applying CART to these data using a minimum node size of 5 on the left and 3 on the right. The cross-validated errors for the 2 trees were 10% and 35%, respectively, with the first being about the same as that for the AIM with 3 markers. The values below each node are the numbers of observations in the training set in each of the 2 classes. We see that in each case, the CART nodes are

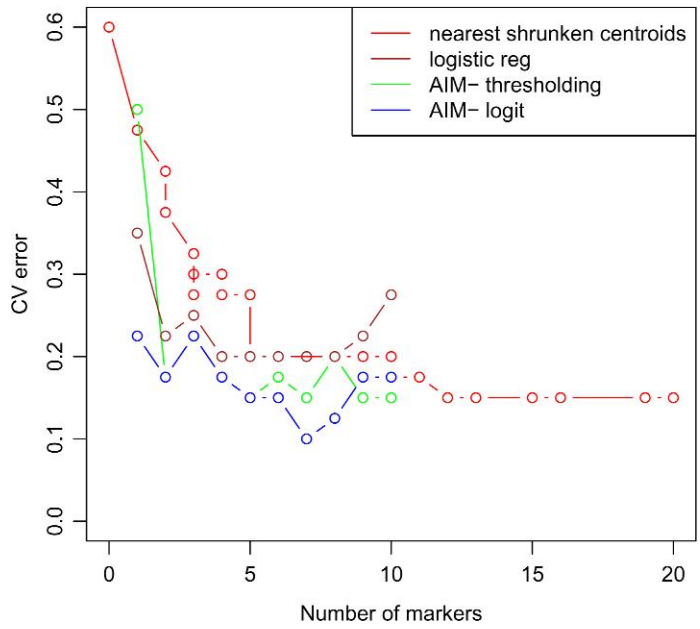

Fig. 3. Protein biomarkers data: cross-validated error curves for AIM and nearest shrunken centroids. The standard error of each curve is about 1%.

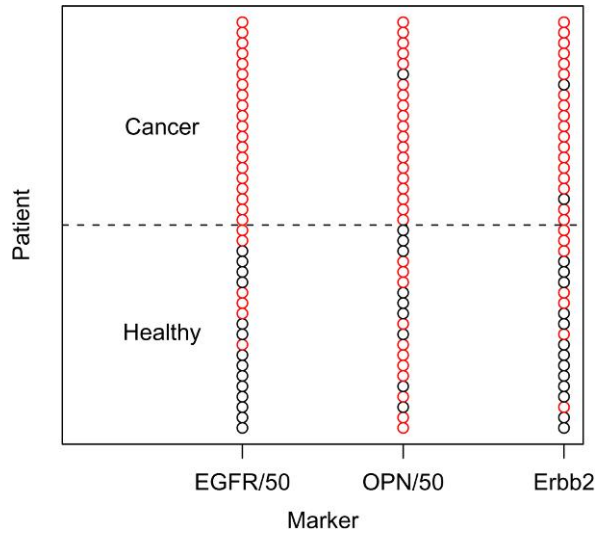

Fig. 4. Protein biomarkers data. Focussing on the model with 3 biomarkers, a red point indicates that the biomarker satisfies its corresponding split condition in that sample.

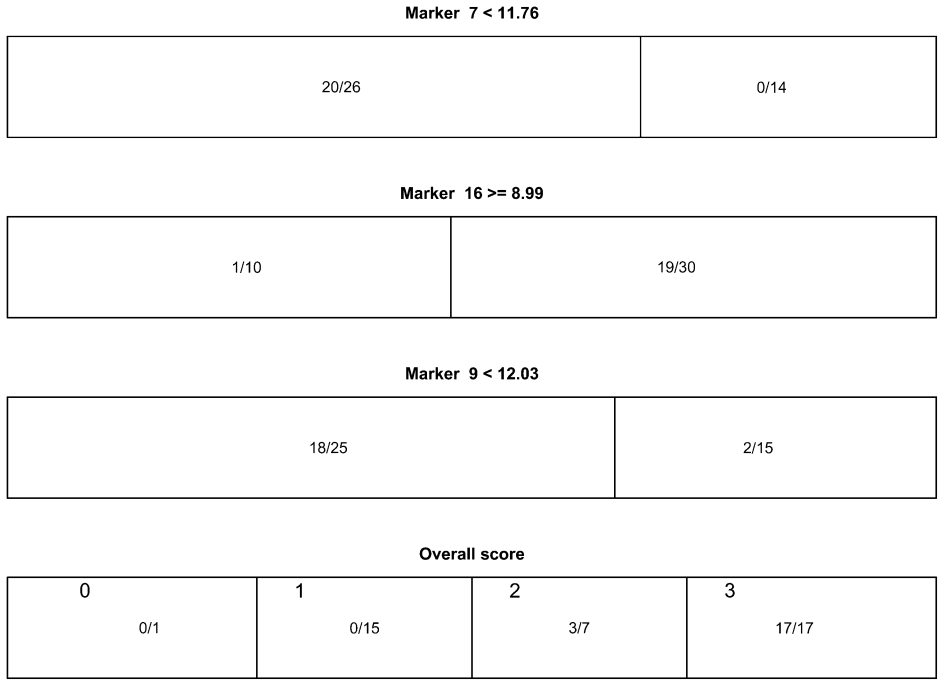

Fig. 5. Protein biomarkers data. Schematic of the 3 splits and final model in bottom panel. The numbers such as 20/26 indicate that there are 20 of 26 patients in the training set with cancer in the corresponding stratum.

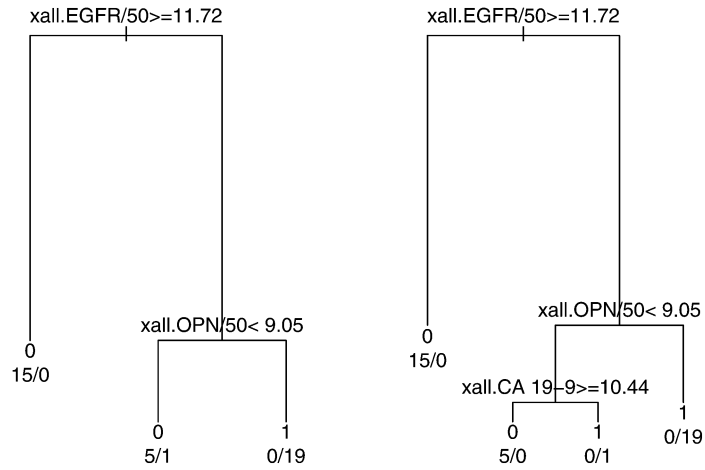

Fig. 6. Protein biomarkers data: CART trees with 3 terminal nodes (left) and 4 terminal nodes (right).

Table 1. *Protein markers example: CV error of GAM using B-splines*

Number of terms	1	2	3	4	5	6	7	8	9	10
CV error	0.56	0.47	0.44	0.58	0.44	0.44	0.36	0.42	0.42	0.50

pure or almost pure. In contrast, the AIM procedure produces a model with a score equal to 0, 1, 2, 3, or 4. The corresponding counts are (1, 0), (15, 0), (4, 3), and (0, 17). Hence, AIM has (potentially) found an intermediate group with a score of 2 and approximately equal numbers of disease and nondisease patients.

Finally, we tried a generalized additive model (GAM) for the purpose of comparison. We allowed a B-spline with 3 degrees of freedom for each marker and used a forward stepwise search. The cross-validated errors are shown in Table 1. With a relatively small sample size, the method seems to overfit fairly quickly and its performance suffers as a result. We also performed the multivariate adaptive spline regression, whose performance is similar to that of GAM in this example.

3. AIM FOR SURVIVAL DATA

In the following section, we present the algorithm for scanning through all possible cutpoints for a given predictor in survival analysis. Here, the outcome $y_i = (T_i, \delta_i)$ and our model is the proportional hazards model

$$h(t|\mathbf{x}) = h_0(t)\exp(\eta), \tag{3.1}$$

where $h(t|\mathbf{x})$ is the hazard function and $\eta = \beta \sum_{k=1}^{K} I(\widetilde{X}_k^* \leqslant c_k)$ for $\widetilde{X}_k^* \in \mathcal{S}$.

Suppose that we want to decide whether to add a binary score $z = (X^* \leqslant c)$ to the existing score s. We may perform the score test as follows. Let $w = s + z$. The score test statistics for testing $\beta = 0$ is $U/V^{1/2}$, where

$$U = \sum_{i=1}^{N} \delta_i \left\{ w_i - \frac{\sum_{l\in R_i} w_l}{n_i} \right\}, \quad V = \sum_{i=1}^{N} \delta_i \left[\frac{1}{n_i} \sum_{l\in R_i} w_l^2 - \left\{ \frac{1}{n_i} \sum_{l\in R_i} w_l \right\}^2 \right],$$

and R_i and n_i are the risk set and its size at time T_i, respectively. Hence, if c moves from x^*_{i-1} to x^*_i, the test statistics can be updated as

$$U \leftarrow U + \delta_i - \sum_{T_k \leqslant T_i} \frac{\delta_k}{n_k}$$

and

$$V \leftarrow V + \sum_{T_k \leqslant T_i} \frac{\delta_k (n_k - 1)}{n_k^2} + 2s_i \sum_{T_k \leqslant T_i} \frac{\delta_k}{n_k} - 2 \sum_{T_k \leqslant T_i} \sum_{T_k \leqslant T_j} \frac{\delta_k \{s_j + I(x^*_j \leqslant x^*_{i-1})\}}{n_k}.$$

3.1 *Lenz data and IPI*

Here, we analyze data on non-Hodgkin's lymphoma from Lenz *and others* (2008). Genetic and clinical information were collected from newly diagnosed lymphoma patients, who received either cyclophosphamide, hydroxydaunorubicin (doxorubicin), oncovin (vincristine), and prednisone/prednisolone (CHOP) or the combination of monoclonal antibody rituximab and CHOP (RCHOP) treatment in 10 institutions at North America and Europe. There are 414 patients, among which 181 patients received CHOP treatment and 233 patients received RCHOP treatment. For the analysis in this section, we will focus on a subgroup of 248 patients, whose IPI is known. The outcome of interest is the overall survival time. Here, we explore whether the AIM procedure can reconstruct the widely used IPI. The details of IPI are given in Section 1. Later, we will analyze the interaction between gene expression and treatment using all the 414 patients.

We divided the data into approximately equal-sized training and test sets and input the 5 predictors (age, stage, LDH, ECOG status, and number of sites) into the AIM procedure. We then computed the Cox score test statistic for resulting index over the test set. The results for a typical training–test split are shown in Figure 7.

This process was repeated 20 times, giving an average score test statistics of 3.87(0.17). The actual IPI had an average score of 3.25(0.14). On the other hand, the average values of Harrell's c-statistic were 0.68(0.01) and 0.64(0.01), respectively. The cutpoints for stage were > 3 versus $\leqslant 3$, 16 of 20 times, and

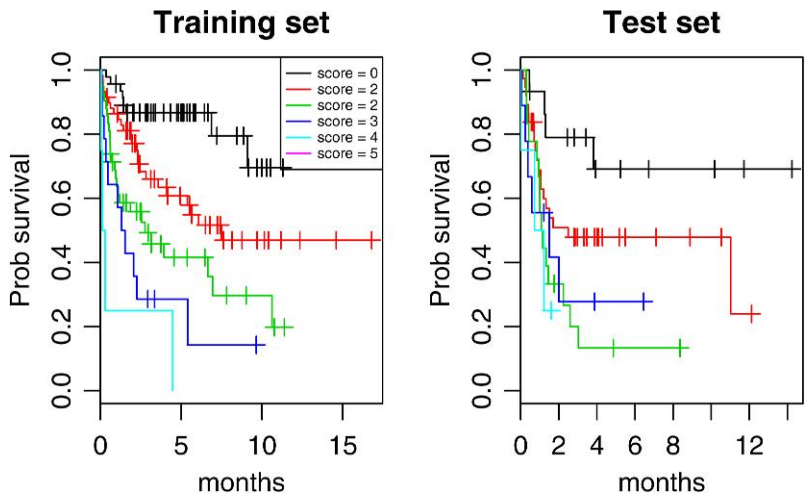

Fig. 7. Survival curves from a set of patients with non-Hodgkin's lymphoma. The patients are stratified by the AIM score estimated from the training set.

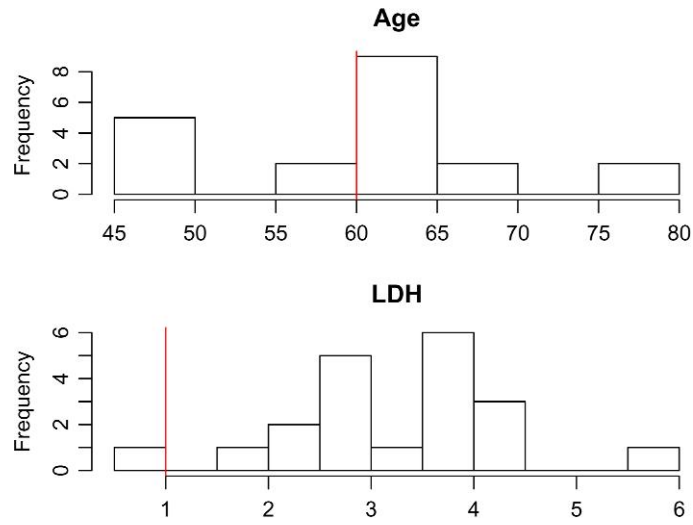

Fig. 8. Lymphoma data: distribution of split points from age and LDH over random splits of the data.

Table 2. *Results for artificial IPI experiment*

Number of noise markers	Average Cox score statistics	Average *c*-statistic	Proportion of markers from the original five
0	3.3 (0.14)	0.64 (0.01)	1.00
2	3.1 (0.21)	0.63 (0.01)	0.69
5	2.5 (0.17)	0.61 (0.01)	0.61
10	2.5 (0.19)	0.61 (0.01)	0.57

> 2 versus $\leqslant 2$, 4 of 20 times, compared to the standard definition of > 2 versus $\leqslant 2$. The numbers of extranodal sites split as > 0 and > 1, 1 and 19 of 20 times, respectively. The distribution of split points for age and LDH are shown in Figure 8 with the corresponding standard IPI split points shown by the red lines. The cutpoints for age are approximately centered at the standard IPI cutpoint of 60, but those for LDH are considerably above the standard IPI cutpoint of 1.0.

Table 2 shows the results when we artificially add standard Gaussian noise markers (independent of the outcome) to the training and test sets. In each case, we used 3-fold cross-validation to choose the model size for AIM. We observe that the procedure still maintains good performance even when a substantial number of noise markers is added.

4. SIMULATION STUDY

In this section, we carry out a small simulation study to investigate the finite sample performance of AIM. In the first set of simulations, we compare the performance of AIM to logistic regression. We generate data in 2 settings, one in which the logistic regression is true and the other in which AIM is true. There are $N = 200$ samples and $p = 10$ predictors, all independent standard Gaussian variables. In each case, only the first few predictors relate to the binary outcome. In the logistic model, $\beta = (1, 1, 2, 2, 3, 0, 0, \ldots)'$

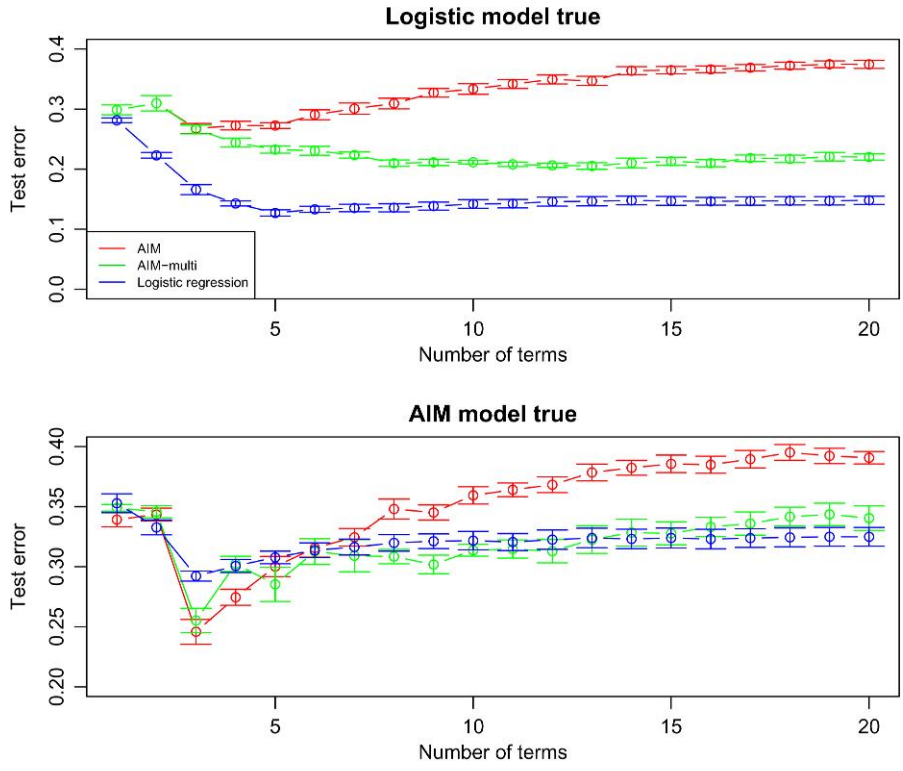

Fig. 9. Simulated data results with $N = 200$, $p = 10$. Mean $\pm$ 1 standard error of the test error for AIM with single and multiple splits per marker and logistic regression. In the left panel, the data were generated from a logistic regression model; in the right panel, the data were generated from the AIM model.

and $\text{Prob}(Y = 1|\mathbf{x}) = \{1 + \exp(-\beta'\mathbf{x})\}^{-1}$. In AIM, $s = \sum_{k=1}^{3} I(X_k > 0)$ and $\text{Prob}(Y = 1|\mathbf{x}) = \{1 + \exp(3 - 2s)\}^{-1}$. Figure 9 shows the results of 10 simulations from these 2 models. It is a sobering reminder that both logistic regression and AIM make modeling assumptions can perform poorly when their underlying assumptions do not hold. In addition, in the bottom panel, AIM shows a tendency to overfit after adding only a few terms. Also included in both panels are the results when AIM is allowed to include up to 5 splits per marker. In the first setting, it performs better than the vanilla AIM procedure, as it tries to approximate the linear effect with multiple binary splits.

In the second set of simulations, we compare the solution obtained by the AIM procedure with the global optimal solution. To this end, we generated data in 2 settings similar as above, one in which the logistic regression is true and the other in which AIM is true. Here, $N = 200$ and $p = 3$. In the logistic model, $\beta = (1, 1, 1)'$ and $\text{Prob}(Y = 1|\mathbf{x}) = \{1 + \exp(-\beta'\mathbf{x})\}^{-1}$. In AIM, $s = \sum_{k=1}^{3} I(X_k > 0)$ and $\text{Prob}(Y = 1|\mathbf{x}) = \{1 + \exp(3 - 2s)\}^{-1}$. For each generated data set, we obtained the AIM solution as well as the optimal solution via exhausted grid search over all possible cutoff points. We then calculated the corresponding z-score test statistics in the generated data set and a independent test set of the same size. Figures 10 and 11 show the results of 500 simulations from AIM and logistic models, respectively. It can be seen that a substantial proportion of the AIM solutions (45% for the AIM setting and 29% for the logistic model) is actually globally optimal and the z-scores from the "suboptimal" AIM solutions are fairly close to that based on the optimal solution in the training set. More importantly, the performance of AIM solution, on average, is comparable to that of the optimal solution in the test set.

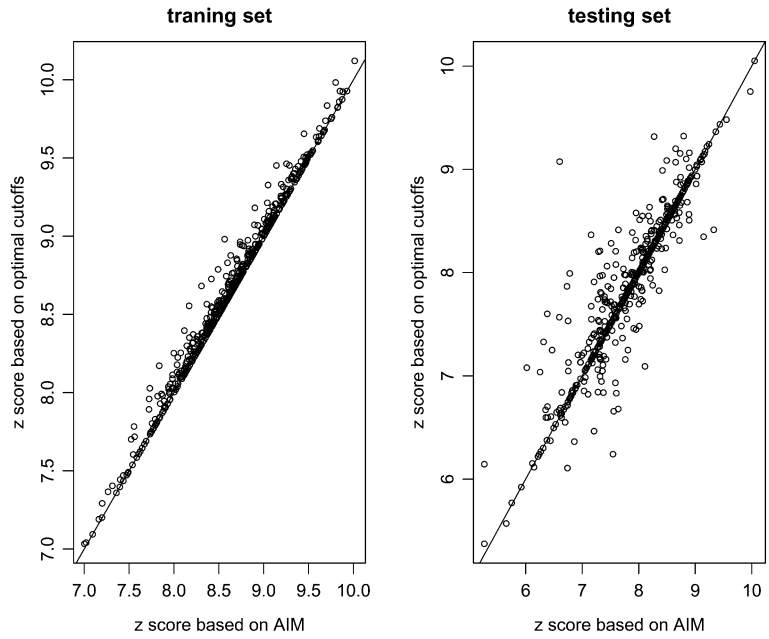

Fig. 10. Simulated data results with AIM. In the left panel, the *z*-scores are based on the training sets; in the right panel, the *z*-scores are based on the testing set.

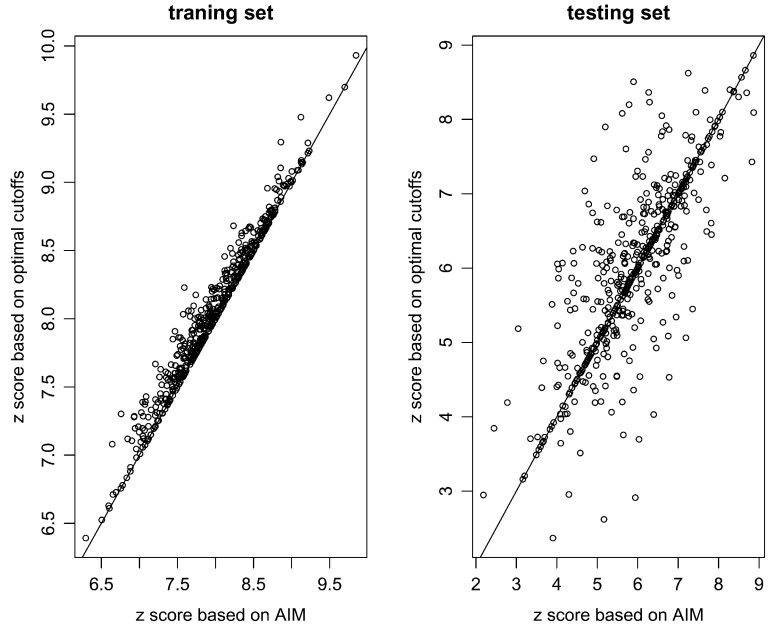

Fig. 11. Simulated data results with logistic regression model. In the left panel, the *z*-scores are based on the training sets; in the right panel, the *z*-scores are based on the testing set.

5. Treatment interactions

In this section, we show how to derive an index that explicitly models the interaction between a set of markers and a binary treatment variable. We will present the algorithm for scanning the cutoff points under linear logistic and Cox models. With the efficient scanning algorithms, the AIM procedure given in Section 2 can be readily used to construct the "IPI"-like index for the interaction of interest. Similarly as above, we assume that s is the existing score and we want to construct the new score in the form of $w = s + I(X^* < c)$.

To determine the cutoff point, we may perform a score test for testing $H_0 : \gamma = 0$ under the assumptions that

$$\mathrm{E}(Y|r, w) = \gamma_0' z + \gamma \cdot (w \times r)$$

and

$$\mathrm{Prob}(Y = 1|r, w) = \frac{\mathrm{e}^{\gamma_0' z + \gamma (w \times r)}}{1 + \mathrm{e}^{\gamma_0' z + \gamma (w \times r)}},$$

for linear and logistic models, respectively, where $z = (1, r)'$ and r is the treatment indicator. We use the logistic model with binary response to illustrate the algorithm. The method for continuous responses is similar. A more general model could also include a main effect for the score w. This would make the computation more complicated but still tractable. However, the interpretation would be more difficult since the contribution of including w to the total likelihood can come from either the main effect or the interaction term. Here, we have chosen instead to search for a pure interaction. Furthermore, in our limited experience, this simple method can still effectively recover the index interacting with the treatment even in the presence of strong main effect.

The score test statistics in logistic regression is $U/V^{1/2}$, where

$$U = \sum_{i=1}^{N} w_i r_i (y_i - \hat{p}_i), \quad V = I_{11} - I_{21} \hat{\Sigma} I_{12},$$

$$I_{11} = \sum_{i=1}^{N} w_i^2 r_i \hat{p}_i (1 - \hat{p}_i),$$

$$I_{21} = I_{12}' = \sum_{i=1}^{N} w_i r_i \begin{pmatrix} 1 \\ r_i \end{pmatrix} \hat{p}_i (1 - \hat{p}_i) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \sum_{i=1}^{N} w_i r_i \hat{p}_i (1 - \hat{p}_i),$$

$$\hat{\Sigma} = \left\{ \sum_{i=1}^{N} z_i z_i' \hat{p}_i (1 - \hat{p}_i) \right\}^{-1},$$

and $\hat{p}_i$ is the empirical mean of responses given the treatment r_i. Let $\hat{\sigma}$ be the sum of all the entries of the 2×2 matrix $\hat{\Sigma}$. V can be simplified as

$$\sum_{i=1}^{N} w_i^2 r_i \hat{p}_i (1 - \hat{p}_i) - \left\{ \sum_{i=1}^{N} w_i r_i \hat{p}_i (1 - \hat{p}_i) \right\}^2 \hat{\sigma}.$$

When the cutoff point c moves from x_{i-1}^* to x_i^*, we have

$$U \leftarrow U + r_i (y_i - \hat{p}_i)$$

and

$$V \leftarrow I_1 - I_2^2 \hat{\sigma},$$

where

$$I_1 \leftarrow I_1 + (2s_i + 1) r_i \hat{p}_i (1 - \hat{p}_i)$$

and

$$I_2 \leftarrow I_2 + r_i \hat{p}_i (1 - \hat{p}_i).$$

For the survival responses, we have $\{(T_i, \delta_i, r_i, w_i), i = 1, \ldots, N\}$. With slight abuse of notations, we consider the score test under the Cox model

$$h(t|r, w) = h_0(t) \exp\{\gamma_0 r + \gamma\, w \times r\}.$$

The score test statistics is $U/V^{1/2}$, where

$$U = \sum_{i=1}^{N} \delta_i \left\{ w_i r_i - \frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} w_l r_l}{\sum_{l \in R_i} e^{\hat{\gamma} r_l}} \right\},$$

$$V = \sum_{i=1}^{N} \delta_i \left[\frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} w_l^2 r_l^2}{\sum_{l \in R_i} e^{\hat{\gamma} r_l}} - \left\{ \frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} w_l r_l}{\sum_{l \in R_i} e^{\hat{\gamma} r_l}} \right\}^2 \right]$$

$$- \left(\sum_{i=1}^{N} \delta_i \left[\frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} w_l r_l^2}{\sum_{l \in R_i} e^{\hat{\gamma} r_l}} - \frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} w_l r_l \sum_{l \in R_i} e^{\hat{\gamma} r_l} r_l}{\left\{ \sum_{l \in R_i} e^{\hat{\gamma} r_l} \right\}^2} \right] \right)^2 V_0,$$

$$V_0 = \left(\sum_{i=1}^{N} \delta_i \left[\frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} r_l^2}{\sum_{l \in R_i} e^{\hat{\gamma} r_l}} - \left\{ \frac{\sum_{l \in R_i} e^{\hat{\gamma} r_l} r_l}{\sum_{l \in R_i} e^{\hat{\gamma} r_l}} \right\}^2 \right] \right)^{-1},$$

and $\hat{\gamma}$ is the maximum partial likelihood estimator for γ_0 under the null model: $h(t|r, w) = h_0(t) e^{\gamma_0 r}$. We introduce the following notations to present the algorithm for scanning all cutoff points:

$$I_1 = \sum_{k=1}^{N} \delta_k \frac{\sum_{l \in R_k} e^{\hat{\gamma} r_l} w_l^2 r_l^2}{\sum_{l \in R_k} e^{\hat{\gamma} r_l}}, \quad I_2 = \sum_{k=1}^{N} \delta_k \left\{ \frac{\sum_{l \in R_k} e^{\hat{\gamma} r_l} w_l r_l}{\sum_{l \in R_k} e^{\hat{\gamma} r_l}} \right\}^2, \quad I_3 = \sum_{k=1}^{N} \delta_k \frac{\sum_{l \in R_k} e^{\hat{\gamma} r_l} w_l r_l^2}{\sum_{l \in R_k} e^{\hat{\gamma} r_l}},$$

and

$$I_4 = \sum_{k=1}^{N} \delta_k \frac{\sum_{l \in R_k} e^{\hat{\gamma} r_l} w_l r_l \sum_{l \in R_k} e^{\hat{\gamma} r_l} r_l}{\left\{ \sum_{l \in R_k} e^{\hat{\gamma} r_l} \right\}^2}.$$

Thus, when c changes from x_{i-1}^* to x_i^*,

$$U \leftarrow U + \delta_i r_i - e^{\hat{\gamma} r_i} r_i \sum_{T_k \leqslant T_i} \frac{\delta_k}{\sum_{l \in R_k} e^{\hat{\gamma} r_l}} \quad \text{and} \quad V \leftarrow (I_1 - I_2) - (I_3 - I_4)^2 V_0,$$

where

$$I_1 \leftarrow I_1 + (2s_i + 1) r_i^2 e^{\beta r_i} \sum_{T_k \leqslant T_i} \frac{\delta_k}{\sum_{l \in R_k} e^{\hat{\gamma} r_l}},$$

$$I_2 \leftarrow I_2 + 2 r_i e^{\beta r_i} \sum_{T_k \leqslant T_i} \sum_{T_k \leqslant T_j} \frac{\delta_k e^{\hat{\gamma} r_j} \{s_j + I(x_j^* \leqslant x_{i-1}^*)\} r_j}{\left\{ \sum_{l \in R_k} e^{\hat{\gamma} r_l} \right\}^2} + r_i^2 e^{2\beta r_i} \sum_{T_k \leqslant T_i} \frac{\delta_k}{\left\{ \sum_{l \in R_k} e^{\hat{\gamma} r_l} \right\}^2},$$

$$I_3 \leftarrow I_3 + r_i^2 e^{\hat{\gamma} r_i} \sum_{T_k \leqslant T_i} \frac{\delta_k}{\sum_{l \in R_k} e^{\hat{\gamma} r_l}}, \quad \text{and} \quad I_4 \leftarrow I_4 + r_i e^{\hat{\gamma} r_i} \sum_{T_k \leqslant T_i} \frac{\delta_k \sum_{l \in R_k} e^{\hat{\gamma} r_l} r_l}{\left\{ \sum_{l \in R_k} e^{\hat{\gamma} r_l} \right\}^2}.$$

5.1 *Lymphoma data*

In this section, we look again at the Lymphoma data of Lenz *and others* (2008). We applied the AIM procedure to look for interactions between treatment (CHOP or RCHOP) and gene expression. Note that the treatment was not randomized in this study, making interpretation of the results more difficult. The genes were first clustered into 149 "metagenes." We split the 414 patients randomly into training and test sets of equal size. Using $K = 3$ markers, AIM produced a score from 0 to 4 with patient counts (0, 49, 121, 4) and (2, 60, 169, 8) in the training and test sets, respectively. Figure 12 shows the survival curves in the test set, stratifying by score (0, 1) versus (2, 3) in the middle and right panels. Specifically, the hazard ratios between RCHOP and CHOP are 0.463 ($p = 0.002$) in the entire cohort, 0.394 in patients with AIM score <2 ($p = 0.0018$) and 0.608 in patients with AIM score $\geqslant 2$ ($p = 0.3$). Therefore, the procedure has identified a subset of patients that may benefit less from RCHOP treatment than the rest of the patients.

The multivariate regression analysis shows that the constructed index has no significant main effect on survival after adjusting for its interaction with treatment. On the other hand, the AIM procedure is also able to construct an index that is highly predictive of survival disregarding treatment. However, given the strong treatment effect in this study, this seems of less interest than the index for treatment interaction.

Table 3 shows the distribution of IPI in the low and high score groups for those patients having an IPI score available (only about half of the patients had missing IPI scores). Since there is no clear difference in IPI across the 2 strata, IPI alone cannot effectively differentiate these 2 types of patients.

5.2 *Surrogate predictors*

When building a model by searching among a sizable number of predictors, there are often alternative models and predictors that fit the data nearly as well as the chosen model. More specifically, in the AIM

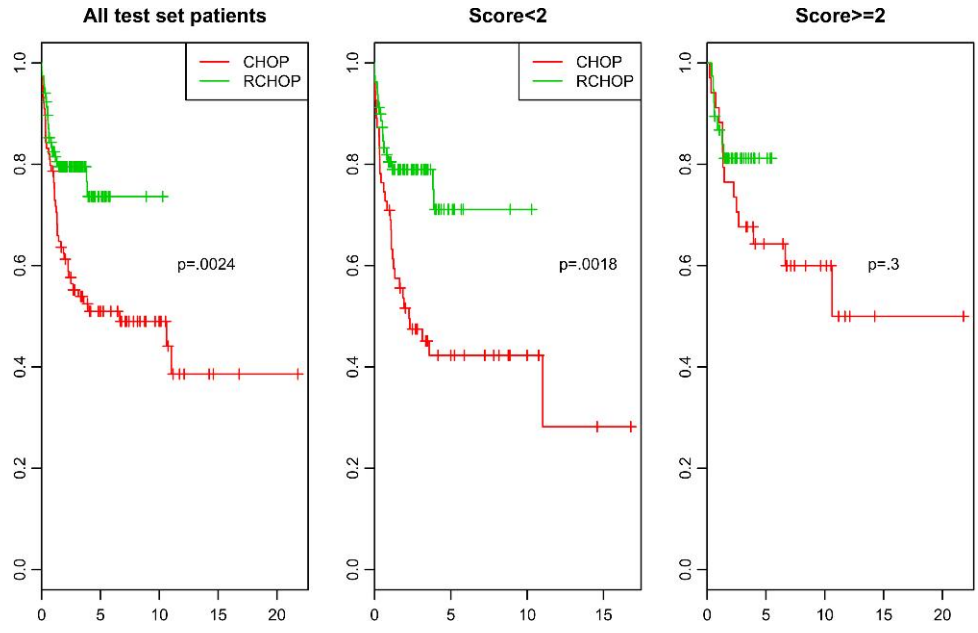

Fig. 12. Lenz data: shown are the survival curves in the test set for all patients in the left panel and stratifying by score (0,1) versus (2,3) in the middle and right panels. The AIM interaction procedure has identified a subset of patients that may not benefit from RCHOP treatment as compared to CHOP.

fit, for any given marker and the corresponding splitting, there may be other markers and split points that produce nearly the same risk stratification. The surrogate markers may be of interest to the scientist and could also be used when applying AIM fit to data in which some of the marker values are not observed for some samples.

In the popular CART procedure, surrogate predictors are found at each split in the tree that best mimic the primary split chosen at that node. These are used to suggest alternative splitting variables and for use when the primary splitting variable is missing in the data.

In the AIM procedure, we find surrogate variables as follows. We apply the one-term logistic regression AIM procedure with the primary marker split as the outcome to find the surrogate marker. Figure 13 shows the best 5 surrogates for the 3 primary markers for the interaction model of Figure 12. The vertical axis shows the misclassification error when using the surrogate marker to predict the primary marker split. We see that each primary marker has at least one surrogate yielding an error rate of about 10%. Figure 14 shows the result of using the top surrogate in place each of the 3 primary markers of Figure 12 and still reveals a potential interaction. Thus, the surrogates are predictive but perhaps not as strongly predictive as the primary markers. As in CART, these surrogates can be used when applying AIM to make predictions in data with missing predictor values.

Table 3. *Distribution of IPI for low and high AIM score groups*

IPI	0	1	2	3	4	5
Score < 2	6	18	20	14	8	0
Score $\geqslant 2$	6	11	17	10	5	2

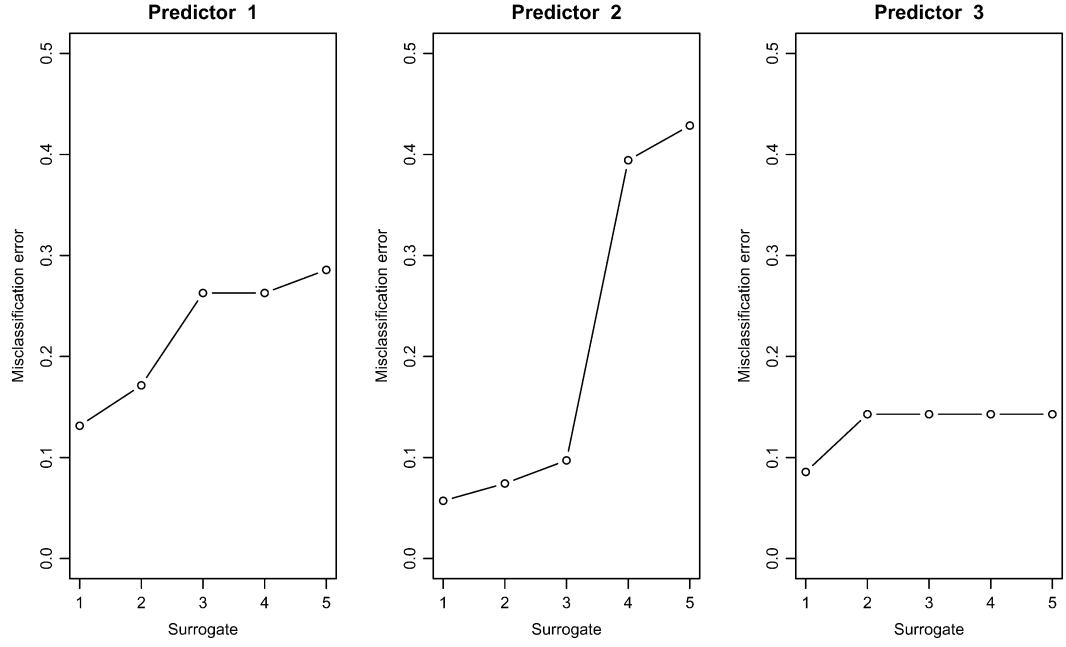

Fig. 13. Lenz data: shown are the best 5 surrogates for the 3 primary markers for the interaction model of Figure 12. The vertical axis shows the misclassification error when using the surrogate marker to predict the primary marker split.

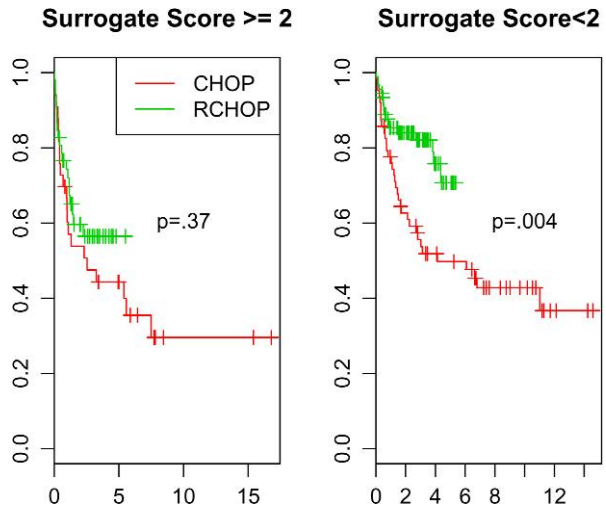

Fig. 14. Lenz data: shown are the survival curves in the test set and stratifying by score (0,1) versus (2,3) in the middle and right panels. Here, we have used the top surrogate marker in place of each primary marker in the score corresponding to that of Figure 12.

6. Degrees of freedom of the AIM fit

In this section, we consider the question: How many degrees of freedom are used in fitting AIM? The degree of freedom measures the intrinsic complexity of underlying models. Beyond purely theoretical interest, the degrees of freedom may have important practical implications: it can help to understand some finite sample behaviors of the AIM procedure, for example why does the prediction accuracy of AIM deteriorate fast after incorporating redundant binary rules in our simulation study; it also can be directly used to select the number of binary rules in AIM via simple criteria such as generalized cross-validation, Akaike information criterion, and Bayesian information criterion. However, as the model is fit adaptively, this is a complicated issue. One popular notion of degrees of freedom is the expected drop in deviance compare to the null model:

$$\mathrm{df}(\hat{\mu}_k) \equiv \mathrm{E}[\mathrm{dev}(\mathbf{y}, \hat{\mu}_0) - \mathrm{dev}(\mathbf{y}, \hat{\mu}_k)], \tag{6.1}$$

where $\hat{\mu}_0$ is the null fit, $\hat{\mu}_k$ is the AIM fit with k terms, and dev is the deviance. A discussion of this definition appears, for example in chapter 7 of Hastie *and others* (2008). For a regression model with k fixed predictors in addition to the intercept, this equals to k. For an AIM fit with k markers, however, this will greatly exceed k for a number of reasons: (i) the split points are found adaptively and (ii) the markers are chosen by an (adaptive) stepwise procedure. With a single predictor, AIM approximately selects the split yielding the maximum drop in deviance, whose exact null distribution has been studied by Worsley (1986) in the context of parametric changing point problem. As $N \to \infty$, it has been shown that the asymptotical null distribution of the maximum drop in deviance converges to that of a maximum of an appropriately scaled Brownian bridge process (Miller and Siegmund, 1982; Combay and Horvath, 1990). More detailed discussions can be found in the Appendix of the supplementary material available at *Biostatistics* online.

In Figure 15, we investigate this numerically. Setting $N = 100$ and $p = 10$, we generate independent standard Gaussian predictors and binary outcomes independent of the predictors. In the left panel, we show fit just X_1 with AIM, ordinary logistic regression, and CART. The logistic regression averages

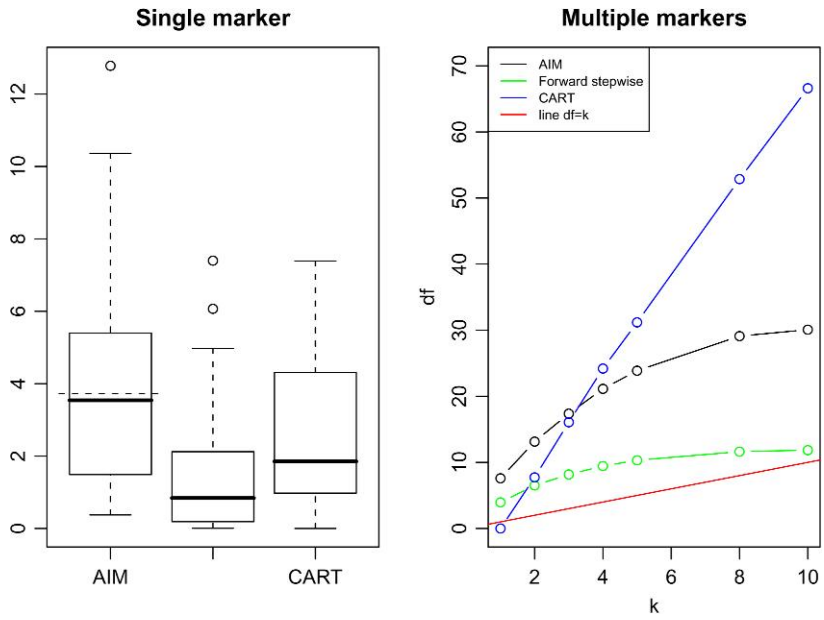

Fig. 15. Estimated degrees of freedom of the AIM, logistic regression, and CART for a single marker (left panel) and multiple markers (right panel). The marker values were standard Gaussian.

about 1 degree of freedom as it should, while CART and AIM average about 2 and 4 degrees of freedom, respectively. On the right panel, we see the result for a k-term model fit by standard forward stepwise logistic regression, AIM, and CART. All exceed k, as we would expect. (For CART, the number of terms refers to the number of splits.) The AIM procedure uses more than 3 degrees of freedom per term near the beginning of the sequence and and fewer than 3 near the end. CART uses more degrees of freedom than AIM as the model grows larger. We also repeat the same experiment, except that each marker was generated randomly from the set of values {1, 2, 3, 4, 5}. The degrees of freedom used by AIM have decreased substantially as compared to Figure 15 since there are fewer possible split points. Some theoretical justifications for the degrees of freedom can be found in the supplementary material available at *Biostatistics* online.

7. DISCUSSION

We have presented a method for adaptive construction of index predictor for regression, classification, and survival analysis.

There is some related work in the literature. The seminal tree–based CART methodology of Breiman *and others* (1984) uses binary splits to produce a decision tree. The terminal nodes (leaves) of the tree are boxes in the feature space. Related to this is the patient rule induction method (PRIM) of Friedman and Fisher (1999), which also constructs boxes in feature space, but they are not connected by a binary tree. LeBlanc *and others* (2005, 2002) refine these methods for survival outcomes and adapt them for clinical trial data. Ruczinski *and others* (2003) introduce logic regression consisting of a set of "and" and "or" rules applied to binary predictors. A simulated annealing procedure is used to estimate the rules.

The AIM methodology is different from and simpler and less ambitious than all these methods. Like CART and PRIM, it makes binary splits on quantitative features. But while those methods combine the

rules with "and," AIM adds them to form a single score. This makes more efficient use of the data in situations where there is a "dose–response" effect involving a set of predictors. For example, given 5 biomarkers $X_1, X_2, \ldots, X_5$, if the outcome risk is proportional $\sum_{k=1}^{5} I(X_k \geqslant c_k)$, then CART or PRIM would have difficulty in capturing this additive structure with a small set of terminal nodes. The single score also facilitates the interpretation of the model produced by AIM fit.

R language software "AIM" is available from CRAN.

SUPPLEMENTARY MATERIAL

Supplementary material is available at http://biostatistics.oxfordjournals.org.

ACKNOWLEDGMENTS

The authors thank the editors and referees for helpful comments that led to improvements in this manuscript. *Conflict of Interest:* None declared.

FUNDING

National Institutes of Health Contract (R01-HL089778 to L.T.); National Science Foundation Grant (DMS-9971405 to R.T.); National Institutes of Health Contract (N01-HV-28183 to R.T.).

REFERENCES

BREIMAN, L., FRIEDMAN, J., OLSHEN, R. AND STONE, C. (1984). *Classification and Regression Trees*. New York: Wadsworth.

COMBAY, E. AND HORVATH, L. (1990). Asymptotic distributions of maximum likelihood tests for change in the mean. *Biometrika* **77**, 411–414.

FREDRIKSSON, S., HORECKA, J., BRUSTUGUN, O. T., SCHLINGEMANN, J., KOONG, A. C., TIBSHIRANI, R. AND DAVIS, R. W. (2008). Multiplexed proximity ligation assays to profile putative plasma biomarkers relevant to pancreatic and ovarian cancer. *Clinical Chemistry* **54**, 582–589.

FRIEDMAN, J. (2001). Greedy function approximation: the gradient boosting machine. *Annals of Statistics* **29**, 1189–1232.

FRIEDMAN, J. AND FISHER, N. (1999). Bump hunting in high dimensional data. *Statistics and Computing* **9**, 123–143.

FRIEDMAN, J., HASTIE, T. AND TIBSHIRANI, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics* **28**, 337–407.

HASTIE, T., TIBSHIRANI, R. AND FRIEDMAN, J. (2008). *The Elements of Statistical Learning; Data Mining, Inference and Prediction*, 2nd edition. New York: Springer.

LEBLANC, M., JACOBSON, J. AND CROWLEY, J. (2002). Partitioning and peeling for constructing patient outcome groups. *Statistical Methods in Medical Research* **11**, 1–28.

LEBLANC, M., MOON, J. AND CROWLEY, J. (2005). Adaptive risk group refinement. *Biometrics* **61**, 370–378.

LENZ, G., WRIGHT, G., DAVE, S., XIAO, W., POWELL, J., ZHAO, H., XU, W., TAN, B., GOLDSCHMIDT, N., IQBAL, J. *and others* (2008). Stromal gene signatures in large-b-cell lymphomas. *New England Journal of Medicine* **359**, 2313–2323.

MILLER, R. AND SIEGMUND, D. (1982). Maximally selected chi-square statistics. *Biometrics* **38**, 1011–1016.

PAUL, D., BAIR, E., HASTIE, T. AND TIBSHIRANI, R. (2008). "Pre-conditioning" for feature selection and regression in high-dimensional problems. *Annals of Statistics* **36**, 1595–1618.

RUCZINSKI, I., KOOPERBERG, C. AND LEBLANC, M. (2003). Logic regression. *Journal of Computational and Graphical Statistics* **12**, 475–511.

TIN-HSLPF, P. (1993). A predictive model for aggressive non-Hodgkin's lymphoma. The international non-Hodgkin's lymphoma prognostic factors project. *New England Journal of Medicine* **329**, 987–994.

WORSLEY, K. (1986). Confidence regions and tests for a change-point in a sequence of exponential family random variable. *Biometrika* **73**, 91–104.

[*Received November 2, 2009; revised June 21, 2010; accepted for publication June 22, 2010*]