Statistics and Computing **15**: 231–239, 2005

Tree-structured subgroup analysis for censored survival data: Validation of computationally inexpensive model selection criteria

ABDISSA NEGASSA*, ANTONIO CIAMPI[†], MICHAL ABRAHAMOWICZ[†,‡], STANLEY SHAPIRO[†] and JEAN-FRANÇOIS BOIVIN[†,§]

**Department of Epidemiology and Population Health, Albert Einstein College of Medicine of Yeshiva University, Bronx NY, USA*
anegassa@aecom.yu.edu
[†]*Department of Epidemiology and Biostatistics, McGill University, Montreal, Canada*
[‡]*Clinical Epidemiology, Montreal General Hospital, Montreal, Canada*
[§]*Center for Clinical Epidemiology and Community studies, The Sir Mortimer B. Davis Jewish General Hospital, Montreal, Canada*

Received March 2003 and accepted February 2005

The performance of computationally inexpensive model selection criteria in the context of tree-structured subgroup analysis is investigated. It is shown through simulation that no single model selection criterion exhibits a uniformly superior performance over a wide range of scenarios. Therefore, a two-stage approach for model selection is proposed and shown to perform satisfactorily. Applied example of subgroup analysis is presented. Problems associated with tree-structured subgroup analysis are discussed and practical solutions are suggested.

Keywords: censored survival data, regression tree, model selection, two-stage approach, subgroup analysis

1. Introduction

Subgroup analysis refers to analysis that is aimed at uncovering possible variation in treatment effect in different patient subgroups such as male/female, young/old etc. The question to be answered by this type of analysis is—for whom does treatment work best? There are varying views regarding the conduct of subgroup analysis (Bulpitt 1988, Gail and Simon 1985, Yusuf *et al.* 1991). However, in the following, we will restrict ourselves to exploratory subgroup analysis.

Ciampi, Negassa and Lou (1995) first presented tree-structured subgroup analysis using the RECursive Partition and Amalgamation (RECPAM) algorithm. The general goal of tree-structured subgroup analysis is to partition patients into groups on the basis of similarity of their response to treatment. The partitioning is based on baseline characteristics such as patient demographics and clinical measurements. In the original presentation by Ciampi, Negassa and Lou (1995), it was assumed that the hazard of patients receiving the new treatment is proportional to the hazard of those receiving the standard treatment. Under this setup, the parameter of interest is the coefficient of the Cox proportional hazards model (Cox 1972) associated with treatment. The final tree structure, selected on the basis of a specific model selection criterion, provides treatment effect within each terminal node, i.e., *subgroup*.

Subgroup tree construction in RECPAM is not based on a formal test of significance of interaction effect between treatment and covariates. Just as in tree-structured analysis for prognostic classification, it is based on a maximally selected statistic as described by various authors including Zhang and Singer (1999), Ciampi *et al.* (1991), Ciampi, Negassa and Lou (1995), LeBlanc and Crowley (1992, 1993), Davis and Anderson (1989), and Segal (1988), i.e., regardless of statistical significance. In the case of prognostic classification, the maximally selected statistic

compares equality of survival experience between the resulting partitions. In contrast, in the case of subgroup analysis, the maximally selected statistic compares whether the *treatment effect* is the same *across* the resulting partitions, i.e., subgroups (Ciampi, Negassa and Lou 1995). In other words, it is based on an optimal value of a criterion statistic for assessing heterogeneity of treatment effect between the resulting subgroups. Nonetheless, a formal test of statistical significance of this heterogeneity would be useful to guard against spurious findings of variation in treatment effect.

In this paper, we will investigate model selection in tree-structured subgroup analysis based on RECPAM. Section 2 deals with performance indicators, and design of simulation. Results of the simulation are presented in Section 3, and an application example is presented in Section 4. Finally, in Section 5, we discuss the implications of our findings.

2. Subgroup analysis tree construction

The observations are assumed to be of the form:

$$(t_i, \delta_i, x_i, \mathbf{z}_i), \quad i = 1, \ldots, N$$

where x denotes treatment, t time-to-event of interest, δ censoring indicator, and $\mathbf{z}$ vector of covariates—potential effect modifiers. The aim of this analysis is to identify, in terms of $\mathbf{z}$, subgroups across which treatment effect varies substantially.

The split criterion or test statistic at a node, i.e., a subset of the entire data that can be partitioned, is the partial likelihood ratio statistic (LRS) based on the Cox partial likelihood (Cox 1972) comparing the model:

$$h(t, x, Q(z)) = \exp\{(\gamma_0 x)[1 - Q(z)] + (\gamma_1 x)Q(z)\}h(t/Q(z)) \tag{2.1}$$

with the simple model:

$$h(t, x, Q(z)) = \exp\{\gamma x\}h(t/Q(z)) \tag{2.2}$$

where $Q(z)$ is indicator of response to a simple question (i.e., requiring a yes/no response) concerning the covariate z. For instance, the question could be " Is $z < c$?" for a continuous covariate z, with $Q(z) = 1$ assigned to a "yes" response. Accordingly, $h(t/Q(z))$ is the subgroup specific baseline hazard rate; subgroups being defined by $Q(z)$.

In equation (2.1), γ_0 and γ_1 are regression coefficients of treatment within each subgroup as defined by $Q(z)$. In contrast, γ in equation (2.2) is an overall regression coefficient of treatment under the assumption of homogeneity. Model (2.2) assumes that the effect of x is the same in the resulting sister nodes, i.e., nodes descending from the same parent node. Notice that both models, (2.1) and (2.2), take into account the potential prognostic effect of $Q(z)$ through allowing the baseline hazard to vary across the resulting subgroups, i.e., through $h(t/Q(z))$. Thus, the partial LRS only measures the amount of information that $Q(z)$ carries about the variation of treatment effect.

It is assumed that the underlying model, generating the data, is a tree structure given by:

$$h(t, x, \mathbf{z}) = \exp\{(\gamma_1 x)I_1(\mathbf{z}) + (\gamma_2 x)I_2(\mathbf{z}) + \cdots + (\gamma_p x)I_p(\mathbf{z})\}h_i(t) \tag{2.3}$$

where $I_i(\mathbf{z})$ is an indicator for the ith terminal node, i.e., a node that cannot be split any further, with a baseline hazard $h_i(t)$. Our interest is in the regression coefficient of log-hazard rate on *treatment*, i.e., γ_i, as derived from the Cox proportional hazards model restricted to the ith subgroup, $i = 1, \ldots, p$. Accordingly, equation (2.3) simplifies to:

$$h(t, x, \mathbf{z}) = \exp(\gamma_i x)h_i(t); \quad i = 1, \ldots, p \tag{2.4}$$

The LRS as a split criterion is computationally expensive. In the case of prognostic classification, we have implemented a closed form estimator at the tree growing stage in order to avoid iterations (Negassa *et al.* 2000). The same closed from estimator was employed in enhancing the computational efficiency of subgroup analysis in RECPAM.

2.1. *Minimally biased tree selection criteria*

Depending on the stopping rule employed, the largest tree could become excessively large to lend itself to a meaningful interpretation. Moreover, the largest tree is prone to overfit bias, i.e., it usually contains spurious splits. In order to deal with these problems, RECPAM has a pruning algorithm similar to Breiman *et al.*'s CART (1984). Pruning is the process of generating a sequence of nested sub-trees; starting with the largest tree by sequentially closing sister nodes, i.e., nodes descending from the same parent node, and ending with the root node. Once the sequence of nested sub-trees is generated, we need to select the minimally biased sub-tree from this sequence. Currently, two general approaches are used for selecting the minimally biased tree: (i) computationally inexpensive criteria, such as, AIC (Akaiki 1974), and (ii) computationally intensive, such as, cross-validation. LeBlanc and Crowley (1993) compared some of the computationally intensive approaches. Negassa *et al.* (2000) reported on the performance of computationally inexpensive model selection approaches in the context of prognostic classification. However, the performance of these computationally inexpensive approaches remains unknown in the case of subgroup analysis.

In this paper, we report on the performance of four model selection approaches in the context of tree-structured subgroup analysis: (i) cross-validation (CV) and (ii) the 1 Standard Error (1SE) rule as described by Breiman *et al.* (1984), (iii) the minimum Akaike Information Criterion (AIC) (Akaike, 1974), and (iv) the elbow approach as described by Ciampi, Negassa and Lou (1995). The elbow approach is equivalent to the tree selection rule proposed by Segal (1988). It consists of choosing a sub-tree in the pruning sequence corresponding to a point beyond which the measure of adequacy starts to change *sharply*. Ciampi, Negassa and Lou (1995) implemented this approach

using AIC as an adequacy measure. The intent is to eliminate residual overfitting, i.e., bias that remains after applying the minimum AIC criterion. Information loss (IL) defined as the LRS comparing a sub-tree with the largest tree in the pruning sequence (Ciampi *et al.* 1991) is another measure of adequacy. IL is a measure of the amount of information lost by dropping part of the largest tree. We think that IL would provide a better graphical means of determining the minimally biased tree because it changes monotonically with the pruning sequence.

We have previously formalized the "*sharp change*" in order to minimize the subjectivity associated with the elbow approach (Negassa *et al.*, 2000). Specifically, we employed consecutive differences in IL (i.e., $\mathrm{IL}_{k+1} - \mathrm{IL}_k$) and their ratios [i.e., $(\mathrm{IL}_{k+1} - \mathrm{IL}_k)/(\mathrm{IL}_k - \mathrm{IL}_{k-1})$], where the subscript k indexes sub-trees in the pruning sequence, as a means of identifying the "*sharp change*." We propose choosing the sub-tree in the pruning sequence that corresponds to the very first point where the local maximum of the ratio of consecutive differences (RCD) coincides with a "non-trivial change" in IL. Based on limited sensitivity analysis, we selected a minimum of 5 as the cut-off for "non-trivial change" in IL (Negassa *et al.*, 2000). RCD is analogous to the criterion proposed by Krzanowski and Lai (1988) for determining the number of groups in a data set using sum of squares clustering.

2.2. *Evaluation of the performance of the various tree selection criteria: Correct recovery, optimism and relative inefficiency*

We employed the following criteria for evaluating performance:

(i) The proportion of correct recovery of the underlying structure, i.e., the "true" data- generating tree. This entails recovering the correct partitioning of the predictor space based on the identification of the terminal nodes of the "true" tree.
(ii) Optimism and relative inefficiency; to be outlined below.

As described earlier, pruning produces a sequence of nested sub-trees, i.e., nested models. In RECPAM, the adequacy of the kth sub-tree in the pruning sequence T_k is measured by the observed information content $I\hat{C}(T_k)$—a measure of the amount of information contained in T_k regarding treatment effect heterogeneity. It is computed as the partial LRS comparing a sub-tree T_k with the root node or the null model. Thus, $I\hat{C}(T_k)$ always increases with tree size, leading to overfit bias.

Following the outline of Efron (1983), we introduced the notion of optimism in evaluating performance. We define optimism as the difference between the observed information content of the largest tree T_0; one in which splitting terminates only when observations within a node are sufficiently similar or a node contains small number of observations, and the information content of the "true" underlying structure T_{true}, i.e., $o\hat{p} = I\hat{C}(T_0) - IC(T_{\text{true}})$. It is a random variable with expectation ω_0. The quantity ω_0 is the average optimism under the knowledge of the correct model. We also introduce $I\hat{C}(T_{\text{sel}})$ as the observed information content of T_{sel} , i.e., a sub-tree selected as minimally biased from the pruning sequence by a specific model selection criterion such as minimum AIC. Likewise the corresponding expected optimism, i.e., the average optimism under applying a given model selection criterion, $\hat{\omega_{sel}}$ is obtained as $E[o\hat{p}_{\text{sel}} = I\hat{C}(T_0) - I\hat{C}(T_{\text{sel}})]$. The discrepancy between $\hat{\omega}_{\text{sel}}$ and ω_0 is the residual bias associated with that specific model selection criterion.

If a model selection criterion performs well then the information content of a tree selected by employing such a criterion will be a good estimator of the information content of the "true" underlying structure. Therefore, mean square error, $\mathrm{MSE} = E[I\hat{C}(T_{\text{sel}}) - IC(T_{\text{true}})]^2$, would be an appropriate criterion to assess how well, on average, $I\hat{C}(T_{\text{sel}})$ estimates $IC(T_{\text{true}})$. A convenient overall summary of performance based on MSE is the relative inefficiency (REL) index (Efron 1983). REL is a measure of performance of model selection criterion relative to two scenarios: (i) selecting the correct tree—best scenario and (ii) selecting the largest tree—worst scenario:

$$\mathrm{REL} = \frac{\mathrm{MSE} - \mathrm{MSE}^{\mathrm{ic}}}{\mathrm{MSE}^{\mathrm{zero}} - \mathrm{MSE}^{\mathrm{ic}}} \tag{2.5}$$

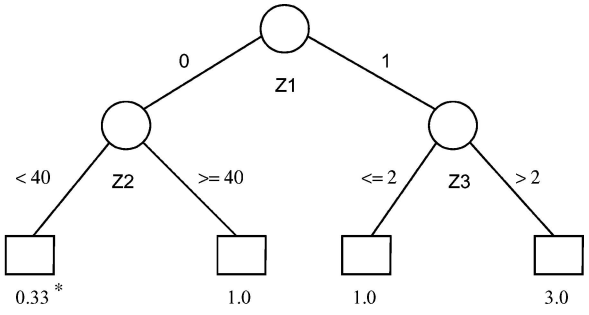
where $\mathrm{MSE}^{\mathrm{zero}}$ is the mean square error of $I\hat{C}(T_0)$ (i.e., obtained under the assumption that $\omega_0 = 0$) and $\mathrm{MSE}^{\mathrm{ic}}$ is the mean square error of the "ideal constant" estimator, $IC^{\mathrm{ic}} = IC(T_0) - \omega_0 \cong IC(T_{\text{true}})$. These are the "worst case" and the "best case" scenarios, respectively, against which performance is assessed. Large REL is indicative of poor performance.

2.3. *Design of simulation*

A total of 600 data sets were generated, with 150 replicates (each with a different random number seed) for each of the four combinations of censoring levels (0 or 50%) and presence/absence of underlying structure. The sample size was fixed at 600.

Data for each individual consisted of survival time, an indicator of censoring, a binary treatment indicator variable and three relevant covariates according to which treatment effect is assumed to vary: one continuous variable, one binary indicator variable and one ordinal variable with five levels. In addition, an individual also has three nuisance covariates (i.e., related to neither treatment nor survival experience): one dichotomous, one ordinal, and one continuous.

Survival times were generated from the exponential distribution. In addition to a complete survival time t, a censoring time c from the same distribution as the survival time was generated. If $\lambda_T(\mathrm{x})$ and λ_C are the hazards for the event time, conditional on treatment, and censoring time distributions, respectively, then the proportion censored in the ith leaf is given by $P_i = 1 - \mathrm{Prob}(t < c) = \lambda_{iC}/\lambda_{iC} + \lambda_{iT}(x)$. We simulated the censored survival time for the kth individual ($k = 1, 2, \ldots, n_i$) in the ith leaf by generating two exponential deviates, T_{ik} with hazard $\lambda_{iT}(\mathrm{x})$ and C_{ik} with hazard λ_{iC}. The response variables for the kth individual are, therefore, the pair (Y_{ik}, δ_{ik}), where $Y_{ik} = \min(T_{ik}, C_{ik})$ and $\delta_{ik} = 1$ if $T_{ik} = Y_{ik}$, and 0 otherwise.



* Relative hazard associated with treatment effect
(Covariates Z4-Z6 are also available but do not modify the treatment effect)

Fig. 1. *Simulation Structure: "True" tree*

The simulation is designed to produce three patient subgroups: (1) those for whom the new treatment is beneficial (RH = 0.33), (2) those for whom the new treatment is harmful (RH = 3.0) and (3) those who do not differentiate between the new and standard treatment (RH = 1). The "true" structure, used to generate the data, is shown in Fig. 1.

3. Results

3.1. *Comparison of the various model selection criteria*

When growing a tree, we employed a stopping rule of a minimum of 50 subjects per node, a minimum of 25 events per node, and a minimum of five subjects per treatment category within a node. As a tree grows larger and larger, leaf-specific treatment effect estimates become increasingly imprecise; stopping rules are employed to avoid this problem.

3.1.1. Simulations with underlying structure

Figure 2 shows results for the case where there is an underlying structure with 0% censoring. The elbow approach exhibited a slightly better performance than cross-validation in identifying the correct structure (56% versus 50%, upper panel). On the other hand, the minimum AIC approach selected structures that were consistently too large while the 1SE rule trees that were too small. Correct recovery by these two criteria was less than 20%. With 50% censoring, the elbow approach performed, again slightly better than cross-validation (38% versus 32% correct recovery, lower panel).

Table 1 presents performance in terms of bias, mean square error and relative inefficiency (REL). Table 1, upper panel, shows that bias associated with cross-validation and the elbow approach was comparable. In contrast, bias associated with minimum AIC and the 1SE rule was substantial. Considering REL as an overall measure of performance, the elbow approach gives the best performance.

3.1.2. Simulations without underlying structure

Under without structure and 0% censoring, the best performance was given by the 1SE rule, followed by cross-validation (Fig. 3, upper panel). The elbow approach tended to select trees with small number of leaves with a mode at the root node, corresponding to the "correct" solution. In contrast, the minimum AIC criterion provided large trees with very low proportion of correct recovery. A similar pattern was observed under 50% censoring (Fig. 3, lower panel).

Consistent with these results, Table 1 (lower panel) indicates best performance by the 1SE rule, then followed by cross-validation. The negative REL associated with the 1SE rule indicates performance superior to the "ideal constant."

3.1.3. Two-stage approach

The above results of our simulation study revealed that there was not a single model selection criterion that was uniformly superior over the range of scenarios considered. The general trend is: (i) the elbow approach gives the best performance whenever there is a structure in the data set, (ii) the 1 SE rule gives the best performance whenever there is no-structure in the data set, and (iii) cross-validation is consistently the *second best*. This observation suggests a two-stage approach to model selection. The first stage involves using a relatively conservative selection criterion to minimize the risk of claiming a structure when there is none, and the second stage involves, provided there is an indication of a structure from first stage, using a reasonably relaxed selection criterion to maximize the chance of identifying the correct structure when there is one.

Under the without structure scenario, the two-stage approach provided the best performance when the 1SE rule was applied at the first stage (see Table 1, end of lower panel). This was to be expected given the results in the previous section.

In the with structure and 0% censoring scenario (see Table 1, end of upper panel), when cross-validation is followed by the elbow approach, the performance of the two-stage approach is exactly the same as using the elbow approach by itself. Given the results in Section 3.1.1, this was expected too. In the case of 50% censoring, this two-stage approach still showed an improvement in terms of reducing bias but suffered from increased variance. This was reflected in the larger REL (see Table 1, end of upper panel) compared to 0% censoring. In contrast, when the 1SE rule is employed at the first stage, there was an improvement in terms of bias reduction. This improvement was markedly counterbalanced by large variation. A similar pattern was observed under 50% censoring.

An intuitive explanation of why the two-stage approach works better is that the two-stage approach reduces inaccuracy (in terms of MSE) without compromising too much on bias. This is best illustrated in Table 1 (upper panel) where the inaccuracy associated with the two-stage approach is smaller than that of CV and its bias is also smaller than or equal to that for the elbow approach. Moreover, in the case of without structure scenario, both inaccuracy and bias associated with the two-stage approach

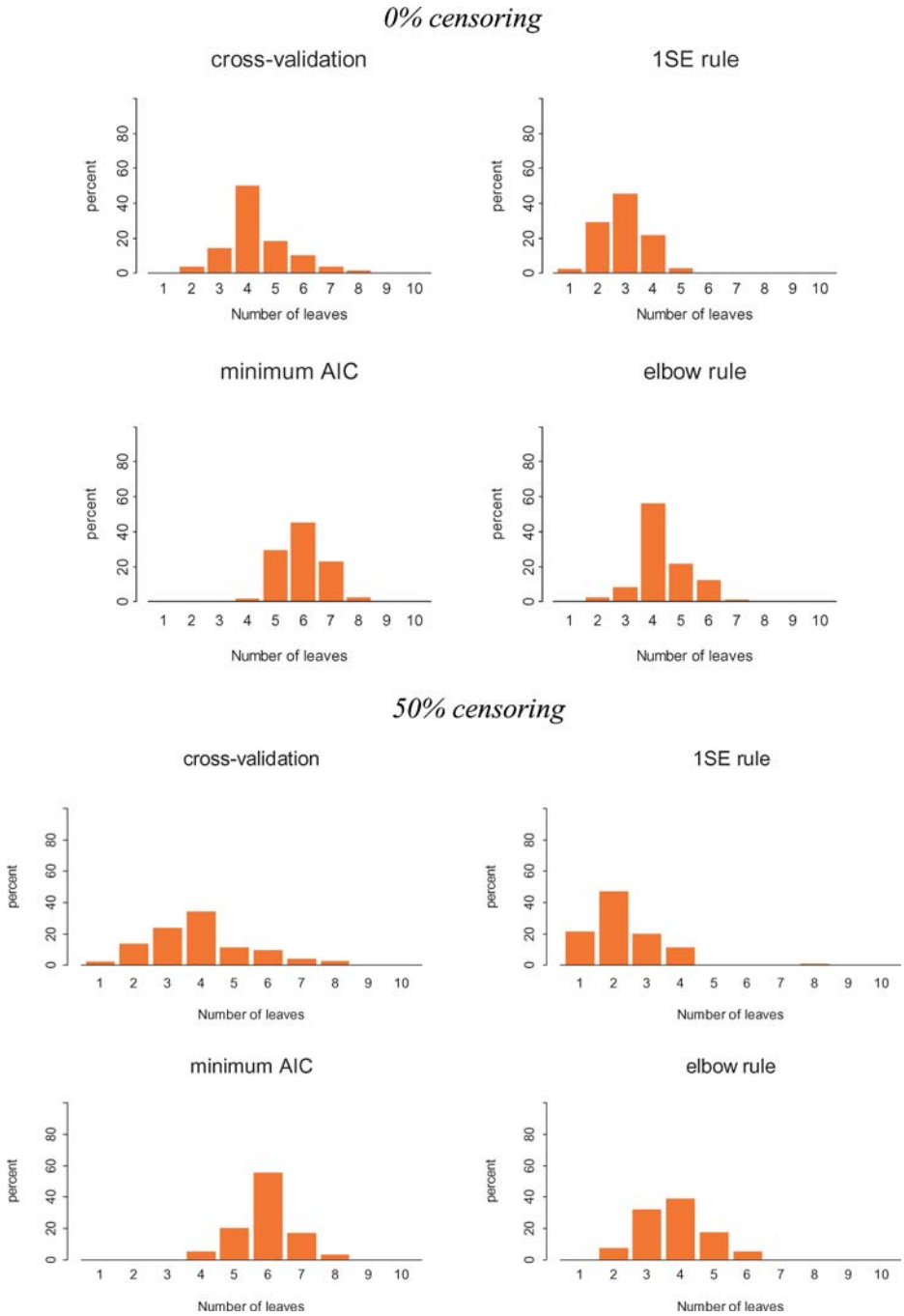


Fig. 2. *Number of leaves by method of tree selection: The true structure has four leaves*

are smaller than that of both CV and the elbow approach alone. In other words, our choice of the two-stage approach is based on a "minimax strategy"; it avoids large inaccuracy (in terms of MSE) of CV in the scenario with structure, as well as higher MSE of the elbow approach when there is no structure (Table 1) and as such minimizes the large errors. Also in terms of correct recovery, our experience suggests that CV does correctly reveal whether there is a structure or not; but tends to prune too much, while the elbow approach reveals the correct structure, if there is one, but is not as good as CV at deciding whether there is a structure (see Figs. 2 and 3). It is this tendency of CV to prune too much, when there is a structure, contributing to its large inaccuracy as pruning an informative split impacts tree adequacy measures substantially as compared to adding/retaining a noise split.

In view of the above results, we recommend the use of cross-validation at the first stage. If cross-validation indicates presence of a structure in the data, i.e., selection of a sub-tree with at least

Table 1. *Bias, mean square error and relative inefficiency by method of model selection*

Method of model selection	Bias		Mean square error (MSE)		Relative inefficiency (REL)	
	0%*	50%	0%	50%	0%	50%
			With structure scenario			
CV	7.07	6.91	336.41	264.52	0.4127	0.3145
Elbow	7.85	7.14	296.94	236.83	0.0444	0.2391
Minimum AIC	14.28	15.8	414.31	395.91	0.6392	0.6724
1SE rule	10.85	9.09	483.71	328.60	0.8410	0.4890
Two-stage with CV	7.85	6.78	296.94	251.89	0.0444	0.4167
Two-stage with 1SE	6.85	1.08	373.55	408.49	0.2151	1.8315
			Without structure scenario			
CV	3.19	2.89	55.91	44.63	0.0571	0.0347
Elbow	6.37	5.33	97.13	77.19	0.1466	0.1161
Minimum AIC	17.04	15.73	330.45	286.44	0.6528	0.6500
1SE rule	0.93	0.91	7.80	9.56	−0.0473	−0.0531
Two-stage with CV	2.61	2.6	43.32	39.40	0.0291	0.0402
Two-stage with 1SE	0.8	1.4	9.0	408.49	0.0001	0.0053

*Level of censoring.

two leaves, then we recommend employing the elbow rule at the second stage to determine the optimal tree size. Otherwise, we recommend stopping at the first stage; concluding absence of structure, i.e., no treatment by covariate interaction.

4. Application to the veteran administration lung cancer trial data set

To illustrate our approach, we analyzed data from the Veteran Administration Lung Cancer Trial (Kalbfleisch and Prentice 1980). The data set consists of 137 subjects, with 6.6% censoring, and eight variables: time-to-event in days, censoring indicator (dead/alive), performance status at randomization, time from diagnosis to randomization (in months), age (in years), prior therapy, treatment (standard/new) and histological type of tumor (squamous, large, small, adenocarcinoma).

In a crude analysis, using Cox proportional hazards model, the treatment effect was not statistically significant $[R\hat{H} = 1.02, 95\%CI(0.71, 1.45), p = 0.93]$. Adjustment for other covariates did not change the result $[R\hat{H} = 1.33, 95\%CI(0.88, 1.99), p = 0.17]$. However, we are interested in exploring the possibility of a treatment by covariate interaction, searching for subgroups of patients for whom the new treatment is better than the standard treatment or *vice versa*.

In growing the tree, we employed a stopping rule of a minimum number of 13 observations per node, a minimum of 7 events per node (i.e., minimum node size ≈10% and minimum event size ≈5% of the total sample size) and a minimum of 5 subjects per treatment category within a node. The largest tree had eight leaves and cross-validation selected the root node. In contrast, the elbow rule selected a sub-tree with three terminal nodes. The patterns of information loss and the ratio of consecutive differences (RCD) in information loss are shown in Fig. 4. The first (from right to left) local maximum of RCD is associated with a moderate change in information loss, 4.69, and corresponds to a sub-tree with five terminal nodes. However, this change is below the cut-point of 5 for an important change (Negassa *et al.*, 2000) and, therefore, it was discarded.

In Fig. 4, the second local maximum of RCD coincided with the first non-trivial jump of 5.2 in information loss. This corresponds to the sub-tree with three terminal nodes. Nevertheless, cross-validation suggested absence of structure in the data set hence our two-stage approach would also lead to the same conclusion.

For illustration purpose, we considered interpretation of the tree identified by the elbow rule. The treatment effects and associated 95% CIs within each subgroup were [4.11, (1.43, 11.80)], [1.39, (0.84, 2.30)], and [0.45, (0.13, 1.61)] in subgroups 1, 2 and 3, respectively. The resulting tree is shown in Fig. 5.

We need to be cautious not to over-interpret the results obtained from this subgroup analysis. This is because these subgroups are not a priori defined and some of the resulting treatment effect estimates are quite imprecise. We carried out a chi-square test for heterogeneity of treatment effect (Schmoor, Ulm and Schumacher 1993) across these three subgroups and found a statistically significant heterogeneity $[\chi^2_{(2)} = 6.99, p = 0.03]$. However, the fact that the groups are identified a posteriori, induces inflation of type I error rate and interpretation requires caution.

Even though the above result seems interesting at face value, it would be more informative to assess the influence of other prognostic factors within each subgroup (i.e., prognostic factors that did not appear in the tree-structure). We adjusted treatment effect for performance status, months from diagnosis and prior therapy within each subgroup. The adjustment changed only the magnitude of treatment effect in the first two subgroups (see Table 2). The change in the regression coefficient of treatment

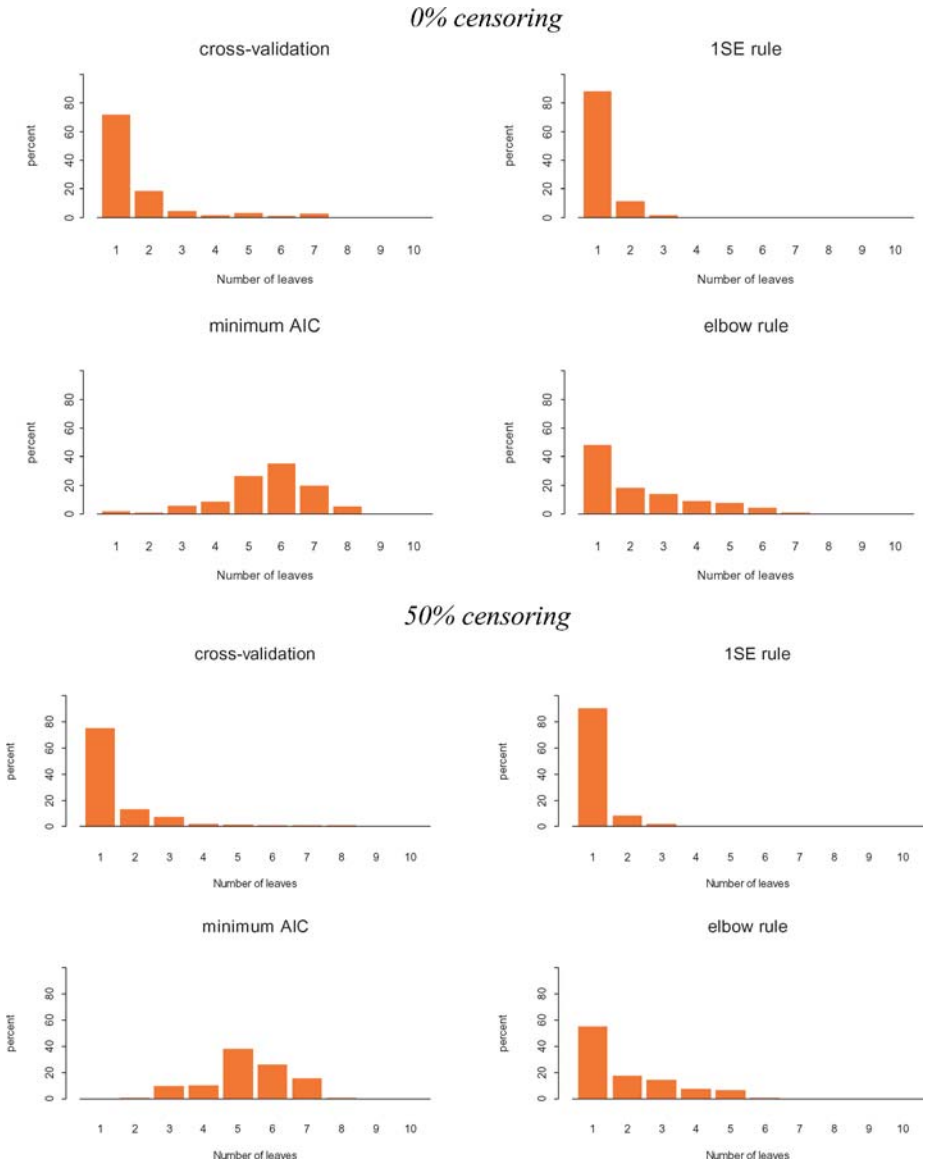


Fig. 3. *Number of leaves by method of tree selection: No structure, i.e., the true structure consists of the root node*

ranged between 5.14–31.37%. Performance status and months from diagnosis were previously identified as important prognostic factors (Ciampi, Negassa and Lou 1995).

After adjustment, there was not evidence for significant heterogeneity [$\chi^2_{(2)} = 3.75$, $p = 0.15$]. This result illustrates the importance of taking into account prognostic factors in conducting subgroup analysis.

5. Discussion

None among the selection criteria evaluated achieves uniform superiority over the range of scenarios considered in this study. Our simulation reveals that the two-stage approach where cross-validation is employed at the first stage, and then is followed by the elbow approach, performs the best. We believe that the two-stage strategy offers a sensible compromise between increasing the chance of identifying the correct structure if there exists one, and minimizing the risk of claiming one when actually there is none. In addition, the formalization of the elbow rule by providing an operational definition of "*sharp change*" reduces the subjectivity associated with graphical assessment and enhances reproducibility of results.

As illustrated by the example, tree-structured subgroup analysis as implemented in RECPAM (Ciampi, Negassa and Lou 1995) has an important limitation that can be easily corrected. At the tree-construction step, while searching for the best split, the effect of other covariates is totally ignored—except for the covariate on which the splitting is being performed. This may result in a spurious variation in treatment effect. Therefore, we suggest that within each resulting subgroup the distributions of

Table 2. *Treatment effect after adjusting for prognostic factors*

Leaf (from left to right in Fig. 5)	$\hat{\beta}(SE)$	p-value	$R\hat{H}$	95% CI
1	0.97 (0.62)	0.116	2.64	(0.79, 8.83)
2	0.41 (0.26)	0.114	1.51	(0.91, 2.52)
3	−0.84 (0.73)	0.247	0.43	(0.10, 1.79)

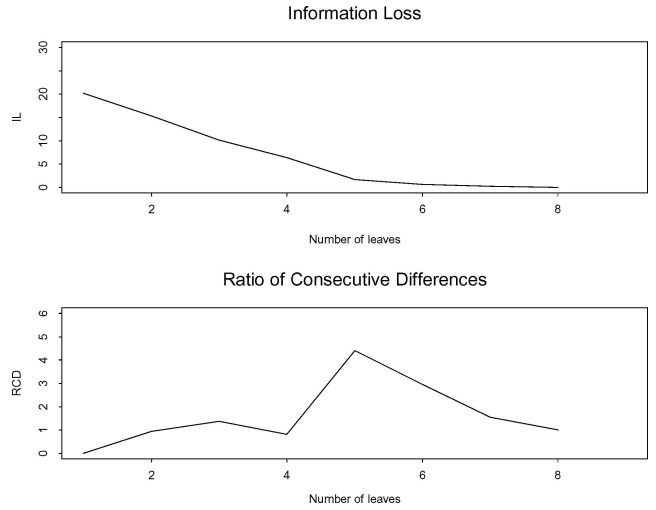


Fig. 4. *Patterns of Information Loss (IL) and Ratio of Consecutive Differences (RCD): Veteran Administration Lung Cancer Data Set*

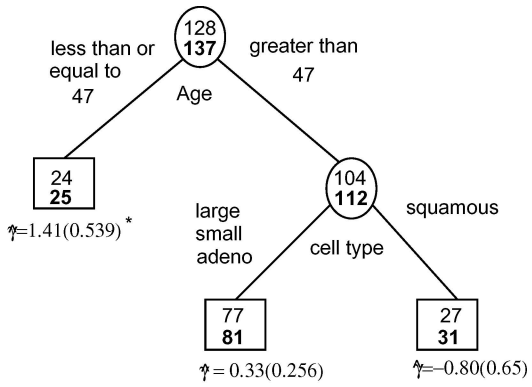


Fig. 5. *Subgroup Analysis: Veteran Administration Lung Cancer Data Set*

important prognostic covariates between treatment categories should be compared and, if necessary, adjusted for. The extent of change in the regression coefficient of treatment after adjustment helps assessing the robustness of the resulting structure.

One of the problems in conducting exploratory subgroup analysis is scarcity of data; negatively impacting the reliability of effect estimates. In most cases, epidemiological studies and clinical trials are planned with enough power to detect only main effects of interest. Therefore, data driven subgroups are likely to be small in size. This may yield wide confidence intervals associated with the subgroup specific estimates of treatment effect, as demonstrated in the case of the veteran administration data. The second issue concerns selecting the covariates for subgroup analysis. This would not be a problem if the investigator has an *a priori* idea as to how to form these subgroups on the basis of clinical experience (this is referred to as "proper subgroups" by Yusuf *et al*. 1991). In this case, there might not be a need for tree-structured analysis, since it would suffice to estimate treatment effect, properly adjusted for potential confounders, within each subgroup, using standard Cox regression. Moreover, if the investigator has a well-formulated hypothesis with respect to a heterogeneous treatment effect, then appropriate statistical power can be ensured at the planning stage. This will also avoid the aforementioned problem of small terminal-node sizes. However, specifying subgroups at the design stage is very rare in practice, and the interest is often in detecting *unexpected* interactions. In such instances, the tree-growing algorithm for subgroup analysis is the appropriate approach and it handles the selection of variables in an "objective" manner, i.e., by maximizing split criterion. The third issue is the role of other important prognostic covariates in the relationship between treatment and outcome. This can be addressed in the same manner as in the case of the veteran administration lung cancer data. Once bias as a possible explanation is ruled out, the structure can be judged with respect to its clinical plausibility, since a chain of clinical statements define the subgroups. Finally, we would like to emphasize that substantively interesting findings of subgroup analysis should be considered as mostly suggestive of the very hypotheses that should be confirmed in an independent data set.

Acknowledgments

This research was supported by the Heart and Stroke foundation of Canada postgraduate scholarship (AN). M.A. is a James McGill Professor at McGill University.

References

Akaike H. 1974. A new look at the statistical model identification. IEEE Transactions on Automatic Control AC-19: 716–723.

Breiman L., Friedman J.H., Olshen R.A. and Stone C.J. 1984. Classification and Regression Trees. Wadsworth, Belmont, CA.

Bulpitt C.J. 1988. Subgroup analysis. Lancet 2: 31–34.

Ciampi A., Lou Z., Lin Q. and Negassa A. 1991. Recursive partition and amalgamation with the exponential family: Theory and application. Applied Stochastic Models and Data Analysis 7: 121–137.

Ciampi A. 1991. Generalized regression trees. Computational Statistic and Data Analysis 12: 57–78.

Ciampi A., Negassa A. and Lou Z. 1995. Tree-structured prediction for censored survival data and the cox model. Journal of Clinical Epidemiology 48: 675–689.

Cox D.R. 1972. Regression models and life tables (with discussion). Journal of the Royal Statistical Society B 34: 187–220.

Davis R.B. and Anderson J.R. 1989. Exponential survival trees. Statistics in Medicine 8: 947–961.

Efron B. 1983. Estimating the error rate of a prediction rule: Improvement on cross-validation. Journal of the American Statistical Association 78: 316–331.

Gail M. and Simon R. 1985. Testing for qualitative interactions between treatment effects and patient subsets. Biometrics 41: 361–372.

Kalbfleisch J.D. and Prentice R.L. 1980. The Statistical Analysis of Failure Time Data. J. Wiley and Sons, New York.

Krzanowski W.J. and Lai Y.T. 1988. A criterion for determining the number of groups in a data set using sum-of-squares clustering. Biometrics 44: 23–34.

LeBlanc M. and Crowley J. 1992. Relative risk tree for survival data. Biometrics 48: 411–425.

LeBlanc M. and Crowley J. 1993. Survival tree by goodness of split. Journal of the American Statistical Association 88: 457–467.

Negassa A., Ciampi A., Abrahamowicz M., Shapiro S. and Boivin J.-F. 2000. Tree-structured prognostic classification for censored survival data: Validation of computationally inexpensive model selection criteria. Journal of Statistical Computation and Simulation 67: 289–318.

Schmoor C., Ulm K. and Schumacher. 1993. Comparison of the cox model and the regression tree procedure in analysing a randomized clinical trial. Statistics in Medicine 12: 2351–2366.

Segal R.M. 1988. Regression tree for censored data. Biometrics 44: 35–47.

Yusuf S., Wittes J., Probstfield J. and Tyroler H.A. 1991. Analysis and interpretation of treatment effects in subgroups of patients in randomized clinical trials. JAMA 266: 93–98.

Zhang H. and Singer B. 1999. Recursive Partitioning in the Health Sciences. Springer-Verlag, New York.