Survival Trees by Goodness of Split

MICHAEL LEBLANC and JOHN CROWLEY*

A tree-based method for censored survival data is developed, based on maximizing the difference in survival between groups of patients represented by nodes in a binary tree. The method includes a pruning algorithm with optimal properties analogous to the classification and regression tree (CART) pruning algorithm. Uniform convergence of the estimates of the conditional cumulative hazard and survival functions is discussed, and an example is given to show the utility of the algorithm for developing prognostic classifications for patients.

KEY WORDS: Classification and regression tree; Regression trees; Survival analysis.

1. INTRODUCTION

Tree-based methods often yield classification and prediction rules that are relatively easy to interpret for a wide variety of applications. Tree-based methods were invented by Morgon and Sonquist (1963). Advances in the practical and theoretical aspects of tree-based methods introduced by Breiman, Friedman, Olshen, and Stone (1984) and the availability of software for their classification and regression tree (CART) algorithm have helped tree-based methods become popular applied statistical tools. Interest in tree-based methods for censored survival data usually comes from the need of clinical researchers to define interpretable prognostic classification rules both for understanding the prognostic structure of data and for designing future clinical trials. In other words, clinical researchers are often interested in forming a small number of groups of patients with differing prognoses. Some applications of tree-based survival techniques were given by Kwak, Halpern, Olshen, and Horning (1990) and by Albain, Crowley, LeBlanc, and Livingston (1990).

In this article we develop a recursive partitioning procedure based on maximizing the dissimilarity in the survival distributions of patients between different regions of the covariate space. Dissimilarity is measured with the logrank test or any other two-sample test useful in the analysis of censored survival data. The logrank test statistic is an attractive measure, because it is in common use for survival analysis and is well studied. The overall performance of the tree structure representing the recursive partition is the sum of the two-sample test statistics between sibling nodes in the binary tree.

The development of recursive partitioning procedures for survival analysis based on the logrank test statistic is not a new idea. But for our tree-based procedure, using between-node dissimilarity, we implement an efficient optimal pruning algorithm equivalent to the CART pruning algorithm and use resampling and permutation techniques to select the tree size.

Segal (1988) developed an algorithm that grows and prunes trees based on the logrank test; however, he does not adopt an explicit measure of performance and an algorithm that gives optimally pruned subtrees, or an "automatic" method of choosing the size of a tree. Ciampi, Thiffault, Nakache, and Asselain (1986) proposed using the Akaike information criterion (AIC) (Akaike 1974) for selecting the tree size, citing the asymptotic equivalence of AIC and cross-validation (Stone 1974). But in this very adaptive setting, involving split point optimization, such an equivalence likely does not hold, because the effective number of degrees of freedom is greater than one per split. Owen (1991) studied similar issues in the context of Friedman's (1991) multivariate adaptive regression spline (MARS) models.

Other tree-based methods for survival data have been proposed that directly adopt most of the CART paradigm (Butler, Gilpin, Gordon, and Olshen 1989; Davis and Anderson 1989; Gordon and Olshen 1985; LeBlanc and Crowley 1992; and Therneau, Grambsh, and Fleming 1990). To extend the CART algorithm, these methods require the choice of a prediction error for censored survival data that both performs well and is intuitively appealing. Butler et al. (1989) also proposed using the logrank test statistic for splitting; however, they used a within-node measure to prune trees and select tree size.

2. NOTATION, SPLITTING STATISTICS, AND TREES

We assume that data include failure time measurements and additional measurements (covariates) that may be associated with failure time. An observation will be distributed as the vector $(T, \delta, \mathbf{X})$, where T is the time under observation, δ is an indicator of failure, and $\mathbf{X}$ is a vector of M covariates. Let U denote the true survival time having cumulative distribution function F, and let V be the true censoring time with cumulative distribution function C. Then assume $\delta = I_{\{U \leq V\}}$, where $I_{\{\cdot\}}$ is the indicator function of the set $\{\cdot\}$, and define the observed time $T = \min(U, V)$. Assume also that U and V are independent given $\mathbf{X}$, for identifiability reasons (Tsiatis 1975). The learning sample consists of the set of iid vectors $\{(T_i, \delta_i, \mathbf{X}_i): i = 1, 2, \ldots, N\}$.

The algorithm splits the data into groups with differing survival. Because the logrank test has been used extensively

* Michael LeBlanc is Assistant Professor, Department of Preventive Medicine and Biostatistics, University of Toronto, Ontario, Canada M5S 1A8. John Crowley is Member, Fred Hutchinson Cancer Research Center, Seattle, WA 98104. This work was supported by Grant P01 GM24472 and, for the first author, in part by a Natural Sciences and Engineering Research Council of Canada postgraduate scholarship when he was a graduate student at the University of Washington. The authors thank Charles Coltman, Robert Livingston, and Kathy Albain of the Southwest Oncology Group for permission to use the lung cancer data described in Section 5 and thank an associate editor and two referees for helpful suggestions.

Journal of the American Statistical Association
June 1993, Vol. 88, No. 422, Theory and Methods

in the analysis of censored survival data, it is a logical choice for measuring dissimilarity in survival between two groups. Let the number of patients in the two groups be n_1 and n_2. Let $Y_1(u)$ and $Y_2(u)$ be the number of individuals at risk in each group at time u; $Y_j(u) = \sum_{k \in R_j} I_{\{T \leq u\}}$, where R_j is the set of observation labels corresponding to group j. Let $\hat{\Lambda}_1(u)$ and $\hat{\Lambda}_2(u)$ be the Nelson (1969) cumulative hazard estimator for each group. The numerator of the logrank statistic can be expressed as a weighted difference between estimated hazard functions,

$$G = \int_0^\infty w(u) \frac{Y_1(u)Y_2(u)}{Y_1(u) + Y_2(u)} (d\hat{\Lambda}_1(u) - d\hat{\Lambda}_2(u)), \quad (1)$$

where $w(\cdot) = 1$. The ratio of this statistic squared divided by an estimate of its variance will be used as a splitting statistic and to define structure. Other weights could be chosen to have greater sensitivity to early or late differences; for example, Harrington and Fleming (1982) weights $w(t) = (1 - \hat{F}(t))^\rho$, where $1 - \hat{F}(t)$ is the estimated survival function at time t and ρ is some fixed positive value. Other measures of structure, such as the Kolmogorov–Smirnov-type statistics for censored survival data, could also be used.

Stability of a statistic under censoring is an important requirement for splitting. If the statistic becomes more variable in censored data, it will tend to split in regions of heavier censoring. The performance of the logrank test statistic for partitioning censored data was investigated in a small simulation experiment; the results presented in Section 4 and others in LeBlanc (1989) indicate that the logrank test performs well for splitting censored data. In addition, there are efficient updating algorithms for logrank test statistics (and permutation variances) for all possible split points. Updating algorithms are easily obtained, because the statistic can also be represented as the linear function

$$G(s) = \sum_{i=1}^{n} I\{X_{ji} \leq s\}(\delta_i - \hat{\Lambda}_0(t_i)),$$

where t_i is time under observation and X_{ji} is the value for covariate j for individual i and s is the split point. This form of the statistic leads to an efficient updating formula for two split points s_1 and s_2,

$$G(s_2) - G(s_1) = \sum_{i=1}^{n} I\{s_1 < X_{ji} \leq s_2\}(\delta_i - \hat{\Lambda}_0(t_i)).$$

It is useful to review some notation and terminology for binary trees (following Breiman et al. 1984) that will be used in this article.

Definition 2.1. A binary tree consists of a finite nonempty set T of positive integers, $1, 2, \ldots, q$, and two functions, left$(\cdot)$ and right$(\cdot)$, from T to $T \cup \{0\}$, which satisfy the following properties for each $h \in T$: (1) left$(h) > h$ and right$(h) =$ left$(h) + 1$ or left$(h) =$ right$(h) = 0$, and (2) for all h in T there is at most one $u \in T$ such that $h =$ right(u) or $h =$ left(u).

Each element of T is called a node. The definitions of mother, daughter, sibling, internal, and terminal nodes are obvious. A root node is a node with no mother node. Figure 1 is an example. In the left frame of Figure 1, nodes 2 and 3 are sibling nodes and left and right daughter nodes of 1; nodes 3, 5, 6, and 7 are terminal nodes. Two other frequently used terms are subtree and a branch.

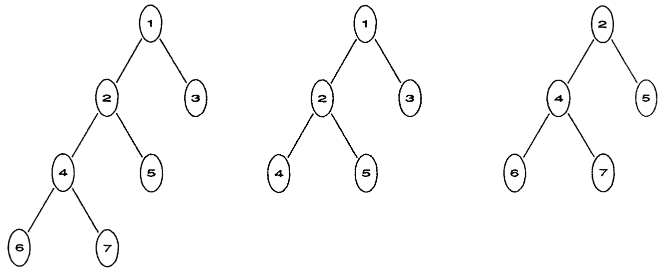

Figure 1. Examples of a Tree, a Pruned Subtree, and a Branch.

Definition 2.2. A tree T_1 is a subtree of T if T_1 is a tree with the same root node as T, and if for every $h \in T_1$, h is in T.

Definition 2.3. A tree T_h is called a branch of T if T_h is a tree with root node $h \in T$ and all descendants of h in T are descendants of h in T_h.

The center and right frames of Figure 1 show a subtree and branch of tree T.

3. THE ALGORITHM

Recursive partitioning algorithms can be described as follows. A rule splits the predictor space $\mathcal{X}$ into two disjoint regions. This rule is applied recursively to the data until the space has been split into many regions, each containing only a few observations. The partition can be represented as a binary tree T, where the set of terminal nodes $\tilde{T}$ corresponds to the partition of the covariate space $\mathcal{X}$ into cardinality $|\tilde{T}|$ disjoint subsets. The components of the algorithm include rules for growing the tree, including the types of partitions that are permitted; rules for pruning the tree back; and rules for choosing a tree of the appropriate size.

3.1 Splitting

A split could be induced by any question of the form "Is $\mathbf{X} \in S$ where $S \subset \mathcal{X}$?" In the CART several forms of splits are possible: splits of a single covariate, splits on linear combinations of predictors, and boolean combination splits.

The simplest class of splits—splits on a single covariate—can be described by the following rules: (1) each split depends on the value of one predictor X_j; (2) if X_j is an ordered variable, then splits induced by questions of the form "Is $X_j \leq c$?" are allowed; and (3) if X_j is nominal with values in $B = \{b_1, b_2, \ldots, b_r\}$ then splits induced by any question of the form "Is $X_j \in S$?" where $S \subset B$ are allowed. Partitioning a node, h, involves finding the split, s, among all variables that maximizes some measure of improvement, $G(s, h)$. In our case $G(s, h)$ is usually a standardized two-sample logrank test statistic. A tree is grown by finding the best split at each terminal node. The best split, s^*, is the split such that

$$G(s^*, h) = \max_{s \in S_h} G(s, h),$$

where S_h is the set of all possible splits of node h. If s^* is not unique, then one of the maximal splits is arbitrarily chosen to partition the data. The same splitting rule is applied recursively to the resulting nodes until a large tree is grown with a small number of observations falling into each node. The tree is large and overfits the data at this point. Hence we describe an algorithm that efficiently finds optimally pruned subtrees using a measure of the tree's performance based on the dissimilarity in survival between sibling nodes in the tree.

3.2 Pruning

3.2.1 Split-Complexity Measure. In CART the cost-complexity measure of tree performance used is

$$R(T) = \sum_{h \in \tilde{T}} R(h) + \alpha |\tilde{T}|,$$

where α is a penalty per terminal node and $R(h)$ is within-node risk. A measure of tree performance analogous to the cost-complexity of CART can be defined for recursive partitions based on two-sample statistics, when within-node error is not used to grow or prune the tree. We call is split-complexity.

Definition 3.1. Define the complexity of T as $|S|$, where $S = T - \tilde{T}$ is the set of internal nodes of T. Call $\alpha \geq 0$ the complexity parameter and define split-complexity $G_\alpha(T)$ as

$$G_\alpha(T) = G(T) - \alpha |S|,$$

where $G(T)$ is the sum over the standardized splitting statistics, $G(h)$, in the tree T:

$$G(T) = \sum_{h \in S} G(h).$$

Call $G(T)$ the goodness of split of a tree T.

One can interpret $G(T)$ as the amount of prognostic structure in the estimate, represented by the tree T. Such an interpretation can be motivated by considering the logrank test statistic as a standardized distance between empirical hazard functions between adjacent nodes in the tree, as displayed in Equation (1). In addition, it is reasonable for an overall measure of prognostic structure to accumulate (or sum) the "distances" between empirical hazard functions among all splits in the tree.

If α is small, then the cost of a large number of splits is small and the tree maximizing $G_\alpha(T)$ will be large. As α increases, the tree maximizing $G_\alpha(T)$ will have fewer nodes, until finally it consists only of the root node.

The split-complexity has the property that if the values of the splitting statistics decrease as one moves down any path in the tree, then the tree that maximizes $G_\alpha(T)$ will be the tree obtained by pruning off all branches that have splitting statistics less than α. Frequently, nonlinearity and interactions can lead to some small values of statistics high in the tree, but these early splits may uncover structure resulting in larger splitting statistics lower down in the tree. Some moderate-sized test statistics can also arise just from noise. This emphasizes the need of some "honest" ways to assess the importance of splits, which we discuss in Section 3.3.

The split-complexity measure provides a way of assessing the trade-off between overall structure found and tree size. Just as with the cost-complexity measure in the CART, there exists a unique subtree that maximizes the split-complexity $G_\alpha(T)$. Consider this the optimal tree for a given α. In the following definition, the symbol "$\preceq$" means "is a subtree of."

Definition 3.2. T_1 is an optimally pruned subtree of T for complexity parameter α if

$$G_\alpha(T_1) = \max_{T' \preceq T} G_\alpha(T'),$$

and it is the smallest optimally pruned subtree if $T_1 \preceq T'$ for every optimally pruned subtree T' of T. Let $T(\alpha)$ denote the smallest optimally pruned subtree of T with respect to α.

The computationally efficient algorithm described in the following section can obtain the nested sequence of optimally pruned subtrees. Evaluating the cost of all possible subtrees would not be feasible even for moderately sized trees, because the number of subtrees grows much faster than the number of terminal nodes in the unpruned tree. To observe how rapidly the number of subtrees grows, consider trees, T^g, for which all terminal nodes have g ancestors. Hence $P(T^0) = 1$, $P(T^1) = 2$, $P(T^2) = 5$, $P(T^3) = 26$, $P(T^4) = 677$, and $P(T^5) = 458{,}330$. Note that T^5 has only 32 terminal nodes. Because trees substantially larger than this may be grown, an efficient algorithm is needed if optimal trees are desired.

3.2.2 The Pruning Algorithm. The pruning algorithm based on split-complexity borrows the idea of weakest link cutting from the cost-complexity algorithm of the CART. For any nonterminal node h of a nontrivial tree T_0, consider the branch T_h. As α increases from 0, there will be a threshold for which the split-complexity for the branch will become less than 0. This value is found by solving the following equality for α:

$$G_\alpha(T_h) = G(T_h) - \alpha |S_h| = 0,$$

where $S_h = T_h - \tilde{T}_h$ is the set of internal nodes of T_h.

Define the function $g(h)$, $h \in T_0$ as

$$g(h) = \frac{G(T_h)}{|S_h|} \quad \text{if } h \in S_0$$
$$= +\infty \quad \text{otherwise,}$$

where $S_0 = T_0 - \tilde{T}_0$.

Then the weakest link $\overline{h_0}$ in T_0 is the node for which

$$g(\overline{h_0}) = \min_{h \in T_0} g(h).$$

Let $\alpha_1 = g(\overline{h_0})$ and let T_1 be the tree obtained by pruning off branch $T_{\overline{h_0}}$. Let $\alpha_2 = g(\overline{h_1}) = \min_{h \in T_1} g(h)$ and let T_2 be the tree obtained by pruning off branch $T_{\overline{h_1}}$. Repeat to obtain the nested sequence of subtrees ($\prec$ means is "a proper subtree of") $T_m \prec \cdots \prec T_k \prec T_{k-1} \cdots \prec T_1 \prec T_0$, where T_m is the root node, and the sequence $\infty > \alpha_m > \cdots > \alpha_k > \alpha_{k-1} > \cdots > \alpha_2 > \alpha_1 > 0$.

Theoretical properties of the pruning algorithm were first derived directly. But although the split-complexity is a measure of dissimilarity between sibling nodes in a binary tree, it can also be expressed as linear function of cost-complexity based on the within node cost, $R(h)$, where

$$R(h) = \sum_{s \in T_{0h} - \tilde{T}_{0h}} G(s) \qquad \text{if } h \in T_0 - \tilde{T}_0$$
$$= 0 \qquad \text{otherwise.} \qquad (2)$$

where T_{0h} is a branch of T_0 rooted at node h. Therefore,

$$G_\alpha(T') = G(T') - \alpha(|T'| - |\tilde{T}'|)$$
$$= G(T_0) - (R(T') + \alpha|\tilde{T}'|) + \alpha$$

for subtrees T' of T_0. Hence any optimality properties of the trees obtained from the pruning algorithm expressed in Theorem 3.1 can be obtained from the properties of the CART algorithm (Breiman et al. 1984, pp. 284–293) and the node cost defined in (2).

Theorem 3.1. The $\{\alpha_k\}$ are an increasing sequence and for $k \geq 1$, $\alpha_k \leq \alpha < \alpha_{k+1}$, $T(\alpha) = T(\alpha_k) = T_k$.

Theorem 3.1 is important given the measure of tree performance because, as in CART, it implies that we can get the best pruned subtree for any penalty α from the efficient pruning algorithm.

3.3 Selection of a Pruned Subtree

In this section we describe some techniques that can be used to select the tree size when only between-node dissimilarity is used to measure tree performance.

3.3.1 Test Sample. The pruning algorithm described in the previous section efficiently yields a sequence of optimally pruned subtrees. We are also interested in selecting one or several subtrees for further exploration. Because the splitting is adaptively based on the survival, the split-complexity $G_\alpha(T)$ is larger than it would be with previously chosen split points; hence the usual distributional results do not hold for the two-sample test statistics. Provided that a large sample is available, this problem can be avoided by dividing the data into a learning sample used to grow the tree, $\mathcal{L}_1$, and a test sample to evaluate the tree's performance, $\mathcal{L}_2$.

The tree is grown and pruned using sample $\mathcal{L}_1$. The nested sequence of trees $T_m \prec \cdots \prec T_k \prec T_{k-1} \cdots \prec T_0$, where T_m is the root node, is obtained using the pruning algorithm described in the previous section. The test sample data is sent down each of these trees. The splitting statistic $G(h)$ is calculated for each internal node h using the validation sample, where the partition is calculated from the learning sample. The tree that maximizes the split-complexity $G_{\alpha_c}(T)$ is chosen as the best pruned subtree, where α_c is the penalty chosen for each split. Because the adaptive choice of split points was made only on the learning sample, we penalize model complexity as we would for less adaptive modeling.

The penalty α_c for each split is typically chosen such that $2 \leq \alpha_c \leq 4$ if the test statistic is approximately χ_1^2 distributed in the two-sample case when there is no difference between groups. A penalty $\alpha_c = 4$ corresponds roughly to the .05 significance level for a split, and $\alpha_c = 2$ is in the spirit of the AIC (Akaike 1974). We emphasize that α_c is not chosen to correct for the very adaptive split point selection, which has been addressed by using a test sample. The penalty (or correction) is needed because even if there is no structure, a logrank statistic based on the test sample is always nonnegative and asymptotically χ_1^2 distributed. Therefore, $G_{\alpha_c}(T)$ with $\alpha_c = 0$ could only increase for more complex trees in the sequence.

3.3.2 Resampling to Correct for Overoptimism Due to Split Point Optimization. Frequently a large validation sample is not available to assess the performance of a tree. A general technique proposed by Efron (1983) for bias correction in the prediction problem could be used to guide in the selection of a subtree.

Let $G(\mathbf{X}_1; \mathbf{X}_2, T) \equiv G(T)$, where $\mathbf{X}_2$ represents the sample used to build T, and the statistic is calculated by sending $\mathbf{X}_1$ through T. Define the following quantities:

$$G^* = E_F G(\mathbf{X}^*; \mathbf{X}, T),$$
$$G = G(\mathbf{X}; \mathbf{X}, T),$$
$$o = G^* - G,$$

and

$$\omega = E_F\{G^* - G\}.$$

If F were known, then $G(\mathbf{X}; \mathbf{X}, T) + \omega$ could be used as a bias-corrected $G(T)$, where ω denotes the overoptimism due to split point optimization. To calculate this quantity in practice, one could replace F (the true distribution of the data) with the empirical distribution of the training sample and use Monte Carlo techniques to estimate ω.

To do this, the tree is grown and pruned on the entire learning sample, and a sequence of nested optimally pruned subtrees is obtained. Set $\alpha'_k = \sqrt{\alpha_k \alpha_{k+1}}$. Remember that T_k is optimal for $\alpha_k \leq \alpha < \alpha_{k+1}$. Draw B bootstrap samples. For each bootstrap sample b, grow a tree, find the optimally pruned subtree for each α'_k, and calculate

$$o_{b_k} = G(\mathbf{X}; \mathbf{X}_b, T_b(\alpha'_k)) - G(\mathbf{X}_b; \mathbf{X}_b, T_b(\alpha'_k)). \qquad (3)$$

Then take the mean over the bootstrap samples,

$$\hat{\omega}_k = \frac{1}{B} \sum_{b=1}^{B} o_{b_k}.$$

Choose the tree that maximizes $\hat{G}_{\alpha_c}(T(\alpha'_k))$, where

$$\hat{G}(T(\alpha'_k)) = G(\mathbf{X}; \mathbf{X}, T(\alpha'_k)) + \hat{\omega}_k.$$

The penalty α_c can be chosen as it was for the test sample case. It is well known that for the prediction problem, in highly overfitted situations, the bootstrap estimate of ω can be quite biased (Efron 1983). An example for trees was given by Breiman et al. (1984, sec. 11.7), showing that the bias can be present even for large sample sizes. Crawford (1987, 1989) studied the bootstrap and other resampling plans for the CART. He showed, through simulation studies, that the ordinary bootstrap can yield considerably biased estimates of the error rate in CART classification. Here this may not be as much of a problem, because for our procedure only the partition is obtained from the training sample. In the prediction problem estimates of the response are also obtained from the training sample. The number of bootstrap samples we typically use are $25 \leq B \leq 100$. For most simulated and real examples considered, there was little difference in sizes of the trees selected for $B \geq 25$. Results of a small simulation study are given in Section 4.2.

3.3.3 Permutation and Approximate Techniques. The performance of a particular division in the tree can be assessed by comparing the observed splitting statistic to an estimate of the permutation distribution of the splitting statistic. Although there are still multiple comparison issues to consider because of the number of nodes (and splits) in a tree, we consider them useful for assessing splitting statistics, given the tree above a given split.

Consider a split at node h. Assume that the censoring and survival distributions do not depend on the measurement values within the node. Therefore, the two tuples $\{(T_i, \delta_i)$: $i \in I_h\}$ are iid, where I_h is the set of labels for node h. Let n_h be the number of observations in node h.

Let $G_0(h)$ be the maximal splitting statistic among all variables and split points in the learning sample for the node. Obtain W_0 samples each of size n_h from the $n_h!$ possible permutations of $\{(T_i, \delta_i)$: $i \in I_h\}$ over the vectors of covariates for individuals $\{x_i$: $i \in I_h\}$. For each permutation k, apply the splitting algorithm to the permuted data and obtain the maximal splitting statistic $G_k(h)$. Under the assumption of no structure, the statistics $G_0(h), G_1(h), \ldots, G_{W_0}(h)$ constitute a sample of size $W_0 + 1$ from the set of all $n_h!$ possible values of the splitting statistic.

Estimated p values for each split in the tree can be calculated directly from the sampled permutation distribution. Let

$$W_1(h) = \sum_{k=1}^{W_0} I_{\{G_k(h) \geq G_0(h)\}}.$$

Then an estimate of the p value is

$$\hat{p}(h) = \frac{W_1(h) + 1}{W_0 + 1}.$$

The multiple testing problem is of concern, because there may be many splits in a tree. But calculating $\hat{p}(h)$ for each split, used alone or in conjunction with split-complexity pruning tree selection techniques, will be useful in data exploration.

The number of permutation samples for each split should be large. Therefore, efficient updating algorithms are much more important in this situation. An indication of the substantial reduction in computing times is given at the end of Section 4.1. But even when updating the algorithms used, the number of samples W_0 may still be limited by available computer time.

One referee has suggested the use of asymptotic approximations to the distribution maximal splitting statistics. Miller and Siegmund (1982) and Halpern (1982) have considered the distribution of maximal χ_1^2 random variables from 2×2 tables derived from splitting on a continuous variable. These results were extended to the logrank test with censored data by Jesperson (1986) when the censoring and survival distributions do not depend on the covariate values within the node. Work of Owen (1991) on assessing degrees of freedom in Friedman's (1991) MARS models could also be extended to survival data. But there are difficulties in directly using these results, because we almost always have multiple covariates, and the covariates are often mixed continuous and categorical and often correlated. Still, such approximations may be useful for assessing splitting statistics, especially if interactive tree-based analysis is a goal. Extensions to the work of Jesperson (1986) are currently being investigated.

4. SIMULATION EXPERIMENTS

We investigated the performance of the logrank splitting statistics and the bootstrap bias correction method through two small simulation experiments.

4.1 Logrank Splitting

The efficacy of the logrank splitting statistic was assessed by studying a single split calculation for a range of survival and censoring distribution configurations.

Failure times were simulated from the H^ρ family (Harrington and Fleming 1982),

$$\begin{aligned} F(t; \psi, \rho) \equiv P(U \leq t; \psi, \rho) &= 1 - (1 + \rho t \psi)^{-(1/\rho)} \\ &\qquad \text{if } \rho > 0 \\ &= 1 - e^{-t\psi} \qquad \text{if } \rho = 0, \end{aligned}$$

and censoring times were simulated from the uniform $(0, \gamma)$ distribution. A single covariate $\mathbf{X}_i = i$: $i = 1, \ldots, 2m$ was considered, and the failure time distribution was F_1 if $i \leq m$ or F_2 if $i > m$. If F_1 and F_2 are different, then a good split would divide the previously simulated observations into two groups of about size m. The survival models for F_1 and F_2 are shown in Table 1. The parameter values were chosen such that the median for the failure time distribution F_1 is approximately .35 and the median for F_2 is approximately .70 for each model with $F_1 \neq F_2$. Hazard ratios between the survival distributions F_1 and F_2 are constant or decreasing and are presented in Figure 2.

Only one sample size is reported here: $2m = 100$. The minimum allowed size of a group resulting from a split was 10 observations. One thousand replications were used in this experiment.

The results are presented in Figure 3 as histograms of the split points on covariate X. The first row of Figure 3 shows that for the constant hazard ratio problem the logrank statistic detects structure well, and that the performance decreases only slightly from the uncensored case to the 20% and 50% censoring cases. As expected, the efficacy of the logrank test is somewhat reduced for groups where the hazard ratios are decreasing, as displayed in rows 2 and 3 of Figure 3.

Table 1. Models for the Logrank Splitting Simulation

		F_1	F_2
	Model	ψ_1, ρ_1	ψ_2, ρ_2
Structure	A_1	2.000, .00	1.000, .00
	A_2	2.184, .25	1.092, .25
	A_3	2.888, 1.00	1.444, 1.00
No structure	A_0	1.000, .00	1.000, .00

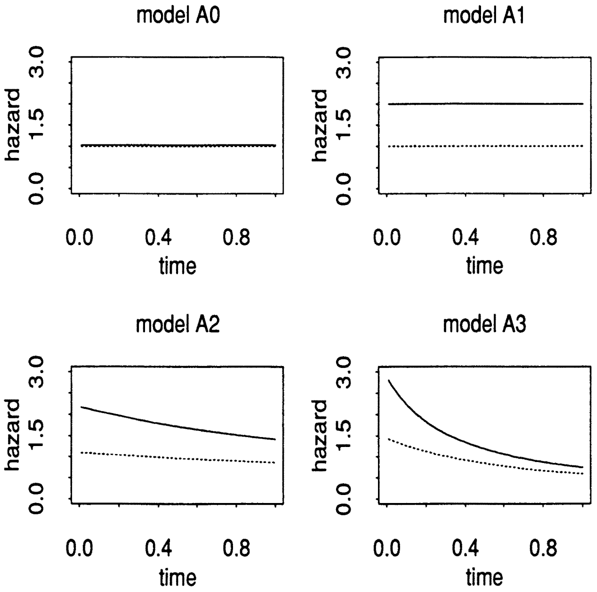

Figure 2. Hazard Functions for Splitting Statistic Simulations. The solid lines and dashed lines represent the hazard functions for F_1 and F_2.

If there is little response structure and the minimum permitted node size is small, then the CART tends to choose splits that send almost all observations to one daughter node. Breiman et al. (1984) called this the end-cut preference. The end-cut preference phenomenon also exists for the logrank test statistic.

Jesperson (1986) proved that for large sample sizes, end-cut preference can also be a problem with the logrank statistic. He considered splitting on a single covariate for the proportional hazards model in large samples under the null and showed that with high probability the split point corresponding to the maximum likelihood ratio, or score test, for any $\varepsilon > 0$ occurs in the ε fraction of the largest or smallest covariate values. But because the minimum number of observations to be split was restricted to be at least 10 (1/10 of the observations) in our simulation, the effect is weak, as displayed in the bottom row of histograms in Figure 3.

The effect of uneven censoring on the splitting statistics was also investigated. For $x \leq 50$ the censoring distribution was $U(0, \gamma_1)$, and for $x > 50$ the censoring distribution was $U(0, \gamma_2)$. The parameters γ_1 and γ_2 were chosen so that there was approximately 20% and 50% censoring in the corresponding regions of the measurement space. Results are displayed only for the null model (bottom right corner of Figure 3); these show that unequal censoring only weakly affects the performance of the logrank splitting.

The simulations were repeated for a smaller sample of size 50 for both the structure case and the no structure case in the survival times. The results were similar to those presented here.

It is also of interest to investigate the effect of some simple updating algorithms on computing time. Clock times on a Silicon Graphics 4D/220 computer with multiple users (so ratios of the times are probably more informative) show that for 1,000 split point optimizations for nodes with 50 observations, computing time was 5 seconds with updating and 43 seconds without updating, and for 1,000 split point optimizations for nodes with 100 observations, computing time was 11 seconds with updating and 133 seconds without updating.

4.2 Bootstrap Bias Correction Method

Five predictor variables $\mathbf{X}_1, \ldots, \mathbf{X}_5$ were generated independently from the $U(0, 1)$ distribution. Survival times were generated from two exponential models (B and C) where the logarithms of the hazard function were

$$B: \quad \theta_i = I\{x_{1i} \leq .5 \cap x_{2i} > .5\}$$

and

$$C: \quad \theta_i = 3.0x_{1i} + 1.0x_{2i},$$

and censoring times were generated from the uniform distribution on $(0, \gamma)$. Uncensored and approximately 50% censored data sets were generated.

We compared the goodness of split procedure with ordinary bootstrap bias correction to a tree-based procedure based on the correct exponential model deviance and that adopts the CART paradigm, including cross-validation to select tree complexity. Such a parametric likelihood tree-based technique seems sensible if the survival model is known; it was studied by Davis and Anderson (1989).

Twenty-five bootstrap samples were used in the split-complexity procedure, and 10-fold cross-validation was used in the exponential CART procedure. We did not include a technique such as 1SE rule of CART. The minimum node size permitted was 20 observations.

The sample size was 250, and 250 samples were generated. We estimated an expected prediction error for the techniques by sending 2,500 observations generated by the same model down the trees and evaluating the correct model deviance.

The simulation yielded quite similar results for the two methods, as shown in Table 2. In fact, for some samples the same tree was chosen by the two methods. But the estimated

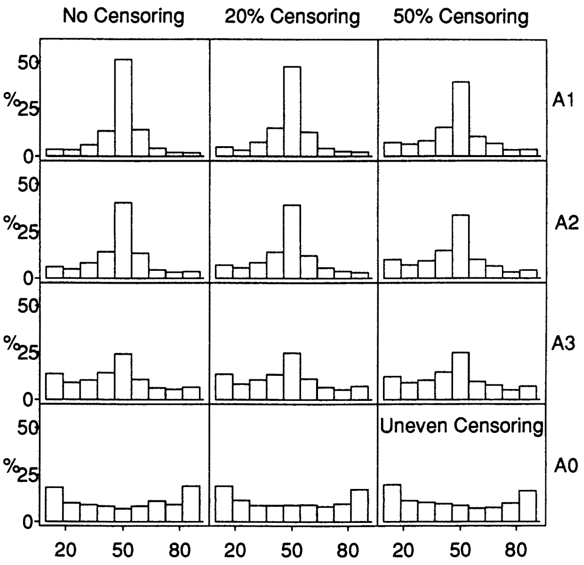

Figure 3. Split Point Frequencies. The percent censoring corresponds to the group with longer lifetimes.

Table 2. Estimated Expected Deviances

	Recursive partitioning technique	
Model	Bootstrap GS	Exponential CART
Expected correct model deviance = 288.6		
B (n = 250)		
No censoring	318.3 (1.50)	314.1 (1.44)
50% censoring	335.6 (1.98)	331.1 (1.70)
C (n = 250)		
No censoring	342.0 (1.24)	344.2 (1.21)
50% censoring	375.9 (2.33)	380.3 (2.13)

exponential deviance for model B for the bootstrap bias corrected method is somewhat larger than that for the other technique. This is likely due to some overfitting. Figure 4 shows a histogram of the tree sizes selected by the two methods. If the procedures correctly identify the split points, for model B, as they did in almost every sample, then the correct tree should have three terminal nodes. Therefore, it is clear that the bootstrap methods can produce trees that are somewhat too large; however, using cross-validation with the correct deviance, which can have high variance, entirely missed the structure in survival times for more than 35% of the samples with approximately 50% censoring. For model C, the split-complexity method out performs the exponential CART method. In other limited simulation studies with nonexponential and even nonconstant hazard ratio models, split-complexity with ordinary bootstrap bias correction performed better than the nonrobust exponential CART, in terms of the expected true model deviance.

Modifications to the bootstrap that reduce the bias problem could be considered. The bootstrap .632 method of Efron (1983), which was shown to perform well for tree-based classification by Crawford (1989), would likely lead to improved performance of the method we propose. The double bootstrap (Efron 1983) is another possibility, although it may be too computationally demanding.

5. CONVERGENCE OF PARTITION-BASED ESTIMATORS

Breiman et al. (1984) showed that as the learning sample becomes large, CART regression estimates converge to the true regression function under certain constraints on the algorithm and smoothness conditions on the regression function. A detailed development of almost sure convergence for tree-based methods was given by Gordon and Olshen (1984). These results can be extended to the case of censored survival data (Butler et al. 1989). Unfortunately, the constraints on the algorithm necessary to show consistency are almost always not implemented. In particular, the algorithm must force the node size or the mesh to go to 0; therefore, the algorithm must split on all variables as the tree grows, not just optimal splits. This is an unattractive modification to our very adaptive algorithm. In addition, we believe that one of the most important applications of tree-based methods is the development of simple, interpretable prognostic stratification rules. To keep the tree interpretable, the size of the tree cannot get too large (or the mesh go to 0) even as the sample size gets large.

Therefore, we emphasize showing uniform convergence of very adaptively chosen partition-based estimators only to the smoothed cumulative hazard and survival functions that are true for recursive partitioning algorithms as usually implemented and that we think are appropriate for our applications. The result follows easily from the convergence results for partition-bsed regression for uncensored data and standard methods used to study the cumulative hazard function and survival function. Convergence is independent of the splitting statistic and pruning algorithm. But it is clear that other statistical properties will depend on good choices for splitting statistics, pruning, and the selection of tree complexity.

To prove consistency, or convergence of Kaplan–Meier estimates to the true conditional survival function, an additional assumption of the mesh size going to 0 is required (Butler et al. 1989; Gordon and Olshen 1980, 1984). Again we restrict ourselves only to convergence to the smoothed hazard functions for nodes; that is, we concern ourselves with the variance term. It would be better if we could theoretically address (not as of yet) the difficult problem of the trade-off between the accuracy of the approximation and tree complexity.

Let the measurement random vector $\mathbf{X} \in \mathbf{R}^M$. This is sufficiently general for the problem where it is assumed that the categorical predictors have at most a uniformly bounded number of classes.

Assume that $\mathcal{B}$ is the collection of all partitions of the measurement space by L linear inequalities into polyhedra. This includes the usual partitions obtained by recursive partitioning, such as the partition of measurement space into boxes

$$B_N = \{(x_1, \ldots, x_M): x_1 \in I_1, \ldots, x_M \in I_M\},$$

where $I_1, \ldots, I_M$ are intervals of the real line. Assume that $B_N \in \mathcal{B}$.

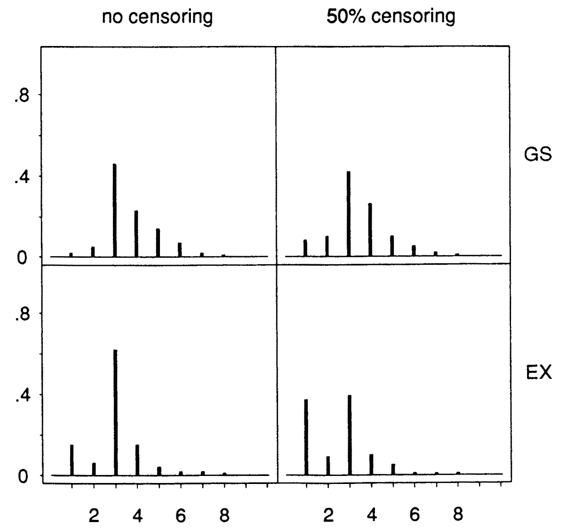

Figure 4. Number of Terminal Nodes for Model B for Trees Chosen by Split-Complexity (Bootstrap Bias Corrected) (GS) and the Tree-Based Method Based on Correct Deviance and Cross-Validation (EX).

Let $v \subset \mathcal{X}$, and define $\eta_N(v)$ to be the number and

$$p_N(v) = \frac{\eta_N(v)}{N}$$

to be the proportion of observations in the learning sample falling into v. Let k_N be nonnegative constants such that

$$P(p_N(v) \geq k_N \frac{\log N}{N} \text{ for } N \geq 1 \text{ and } v \subset \mathcal{X}) = 1.$$

Consider two subdistribution functions used by Breslow and Crowley (1974) to study the Kaplan–Meier estimator. Define

$$1 - H_x(t) = P_x(T > t) = (1 - F_x(t))(1 - C_x(t))$$

and

$$H_{1x}(t) = P_x(T \leq t, \delta = 1) = \int_0^t (1 - C_x^-) dF_x$$

for $0 \leq t < \infty$ and where the censoring and survival distributions are continuous and/or discrete. We consider estimates on the finite time interval $[0, \tau]$ and assume that $H_x(\tau) > 0$. The empirical distribution functions estimates of H and H_1 for $v \in \mathcal{B}$ are

$$H_N(t, v) = \frac{1}{\eta_N(v)} \sum_{\mathbf{X}_i \in v} I_{\{T_i \leq t\}}$$

and

$$H_{1N}(t, v) = \frac{1}{\eta_N(v)} \sum_{\mathbf{X}_i \in v} \delta_i I_{\{T_i \leq t\}}.$$

Smoothed distribution functions are

$$\bar{H}_N(t, v) = \frac{1}{\eta_N(v)} \sum_{\mathbf{X}_i \in v} E[I_{\{T_i \leq t\}} \mid \mathbf{X}_i]$$

and

$$\bar{H}_{1N}(t, v) = \frac{1}{\eta_N(v)} \sum_{\mathbf{X}_i \in v} E[\delta_i I_{\{T_i \leq t\}} \mid \mathbf{X}_i],$$

where expectations are over the distribution of (T, δ) given the covariate. The cumulative hazard estimator for $v \in \mathcal{B}$ is

$$\hat{\Lambda}(t, v) = \int_0^t \frac{dH_{1N}(s, v)}{1 - H_N(s^-, v)},$$

and a smoothed cumulative hazard is

$$\bar{\Lambda}(t, v) = \int_0^t \frac{d\bar{H}_{1N}(s, v)}{1 - \bar{H}_N(s^-, v)}.$$

Theorem 5.1. Assume that $\lim_{N \to \infty} k_N = \infty$; then for every polyhedron B with a uniformly bounded number of faces in $\mathbf{R}^M$ and every $\varepsilon > 0$ and $c > 0$,

$$\lim_N N^c P\Bigl(\sup_{t \leq \tau} |\hat{\Lambda}_N(t, v) - \bar{\Lambda}_N(t, v)| > \varepsilon \text{ for some } v \in \mathcal{B} \text{ such that } v \subset B \text{ and } p_N(v) \geq k_N \frac{\log N}{N}\Bigr) = 0.$$

Convergence can be extended to the Kaplan–Meier estimator by following the argument of Shorack and Wellner (1986, p. 305).

6. EXAMPLE

This section studies a data set based on small-cell lung cancer patients from several recent clinical trials of the Southwest Oncology Group. The main interest of this analysis is to determine whether there is evidence to support a refinement of the current staging system (LD, tumor confined to one hemithorax) and extensive stage disease (ED, distant metastases) staging system to better define prognostic subgroups. Other discussion and examples of tree-based methods for developing strata can be found in Ciampi et al. (1986), Fries et al. (1988), and Bloch and Segal (1989).

The data set consisted of a sample of 704 patients entered on four limited disease trials and two extensive disease trials between 1980 and 1988. Approximately 14% of the patients had censored observations. The following pretreatment characteristics were available for all patients in the sample: sex, age, smoking status (never, stopped >1 year, stopped <1 year, current), alkaline phosphatase, T and N staging status if limited disease, single versus multiple metastases if extensive disease, presence of pleural effusions, performance status (ambulatory or nonambulatory) and lactic acid dehydrogenase (LDH ≤ upper limit of normal versus >upper limit of normal). A complete description of the study design and data were given in Albain et al. (1990).

A tree was grown with the minimum node size set to 25 patients, then the split complexity pruning algorithm was used. The corrected split-complexity based on 25 bootstrap replications was calculated, using a penalty of two per split. A pruned tree with seven terminal nodes is presented in Figure 5, and the corrected split-complexity statistics for each

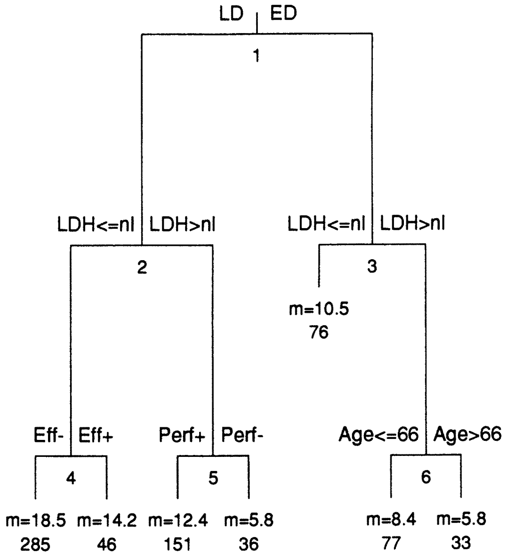

Figure 5. Lung Cancer Example—Pruned Survival Tree. Median-survival (m) in months and the number of observations are given below each terminal node. The variables are disease stage (limited, LD, or extensive, ED); lactic acid dehydrogenase, LDH (≤ normal level (nl) or >nl); pleural effusions, Eff (present (+) or absent (−)); age; and performance status, Perf (ambulatory (+) or nonambulatory (−)).

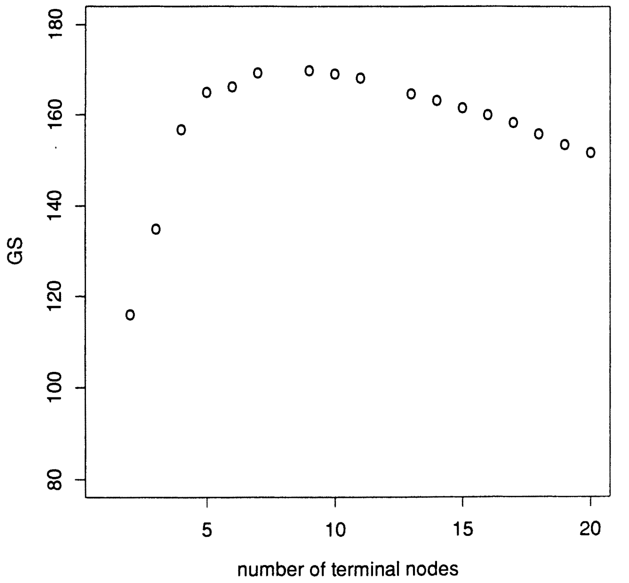

Figure 6. Corrected Split-Complexity (α_c = 2) Versus the Number of Terminal Nodes.

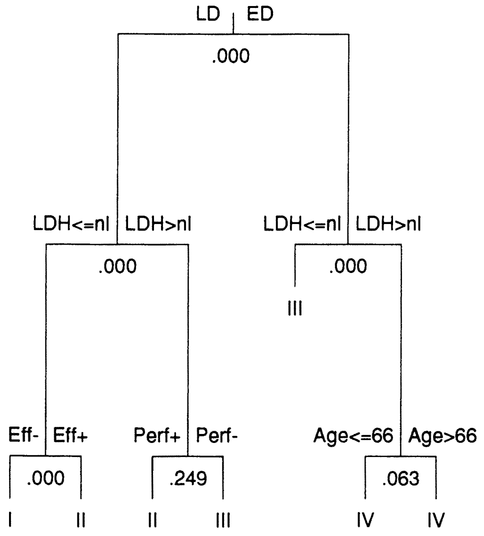

Figure 7. Lung Cancer Example—A Prognostic Stratification. Permutation test p values based on 1,000 permutations are given below each split. The median survival for stages I, II, III, and IV are 18.5, 12.9, 10.3, and 6.5 months.

tree in the pruned sequence are presented in Figure 6. The nine-node tree had very slightly larger split-complexity, but parsimony suggested the next smaller tree in the sequence. In addition, for penalties α_c = 2.5, 3.0, 3.5, and 4.0, the optimal tree was again the seven-node tree presented. Note that the results for this example are not that sensitive to different penalties within the typical range, of 2 to 4.

The first split in the tree was on the extent of disease. The next split for both limited and extensive disease groups was on LDH. The directions of effects were consistent with those seen using Cox's (1972) linear proportional hazards model. The value of the maximal logrank test statistic for each split is given in Table 3. To further assess the prognostic structure and the stability of the tree, the second-largest maximal split statistics are also given in Table 3. There are no strong competitor split variables for the tree.

Often medical researchers wish to use only a small number of stages in designing clinical trials. The number of groups defined by a tree can be reduced by the combination of nodes with similar survival characteristics. This would typically use the investigators' knowledge of the biology; however, an automatic technique to reduce the number of groups would be useful.

In this analysis the terminal nodes shown in Figure 5 were ordered based on ratios of observed to expected deaths in the node based on the null model. Median survival (which in this example results in a slightly different ordering) or some other measure of prognosis in a node could also be used. The monotone ordering was coded as a single ordered covariate, and the recursive partitioning algorithm was used again. Using a penalty of 2 per split resulted in four stages. One possible staging scheme is presented in Figure 7, and Kaplan–Meier estimates are presented in Figure 8. The median survival for stages I, II, III, and IV are 18.5, 12.9, 10.3, and 6.5 months.

Table 3. Splits and First Competitor Splits

Split label	*Chosen split variable (logrank stat)*	*First competitor split variable (logrank stat)*
1	LD/ED (114.9)	LDH ≤ nl/LDH > nl (58.1)
2	LDH ≤ nl/LDH > nl (23.0)	Perf+/Perf− (10.3)
3	LDH ≤ nl/LDH > nl (25.7)	Age ≤ 68/Age > 68 (13.1)
4	Eff−/Eff+ (11.7)	Sex M/F (4.8)
5	Perf+/Perf− (6.7)	Sex M/F (2.1)
6	Age ≤ 66/Age > 66 (7.0)	Perf+/Perf− (1.9)

The approximate significance of splits given the structure higher in the tree was also calculated. The permutation distribution of the splitting statistic can be estimated under the assumption that the censoring and survival times do not depend on the measurement variables in the node. The es-

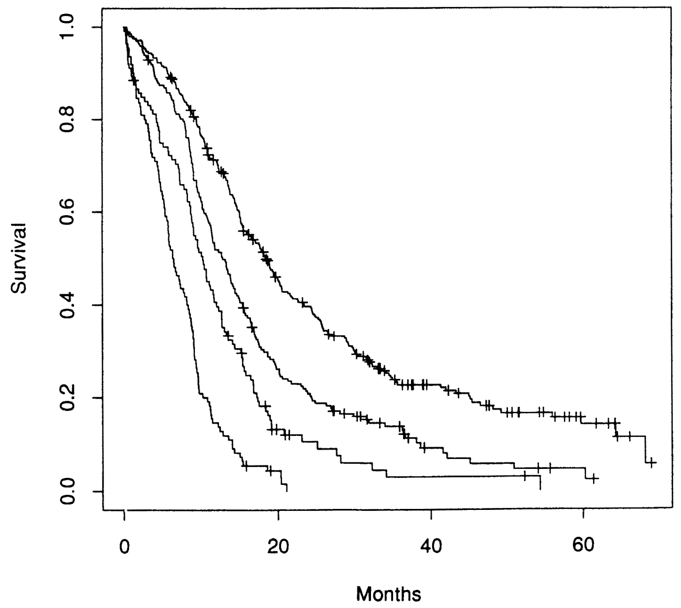

Figure 8. Lung Cancer Example—Survival by Stage (I, II, III, and IV).

timated p values, based on samples of 1,000 from the permutation distribution, are presented in Figure 7. The p value (.249) corresponding to the split on performance status indicates that one may not want to retain that particular split in the tree.

Based on this brief analysis, there is evidence to support the refinement of the current extensive disease and limited disease staging system. LDH is an important prognostic factor that could be useful in staging schemes for small-cell lung cancer. For a more complete analysis based on a larger number of patients, see Albain et al. (1990).

Modifications could be considered to force stages to represent a substantial proportion of the patient population. This could be implemented in the second-stage recursive partitioning procedure by specifying a minimum node size; for instance, 10% of the observations in the learning sample. If more equal-sized groups was a greater concern, the nodes could be ordered as was previously described and assigned to a fixed number of J stages that divide up the observations in the sample as evenly as possible. There are other possibilities for amalgamation of nodes. Ciampi, Hogg, McKinney, and Thiffault (1988) suggested calculating all pairwise two-sample logrank test statistics between nodes. The two nodes corresponding to the smallest value of the statistics are combined, and the procedure is repeated.

7. DISCUSSION

The goal of this article was to further develop a tree-based methodology for the analysis of censored survival data based on the notion of dissimilarity in survival distributions between groups of patients. Where possible, we have tried to use the familiar and well-tested engineering of the CART algorithm. Gordon and Olshen (1985) first extended the CART algorithm to censored data using a within-node measure of fit. For our method based on two-sample test statistics, we have retained an optimal pruning algorithm of the CART and resampling methods to select tree complexity. In addition, we have discussed asymptotics of procedures for censored data that partition the covariate space into a small number of prognostic groups. Results of Crawford (1989) and a small simulation study presented here suggest that the improvements to the bootstrap method we propose for choosing tree size could also be obtained by using the bootstrap .632 method of Efron (1983).

The small-cell lung cancer example focused on the development of prognostic stratification or staging schemes. But that is not the only practical use for the methodology. The simple framework provides a flexible method for the general exploration of censored survival data. This is particularly true if the method were used in conjunction with other simple tools, such as scatterplots and estimates of the survival function. These other tools can greatly aid interpretation of the pruned trees and the structure of survival. For instance, scatterplots of variables involved in splits and other variables are important for better understanding the relationships expressed in the data. Plots of the Kaplan–Meier estimates of the survival distributions are useful descriptions of the prognosis within nodes of the tree. The bootstrap trees derived in the model selection procedure can also be used to assess tree stability.

Implementation of recursive partitioning techniques in good environments for statistical analysis is particularly important. The S computing language (Becker, Chambers, and Wilks 1988) is currently one of the most-developed interactive programming environments for data analysis and graphics. S functions based on tree-based tools of Clark and Pregibon (1991), which allow the investigator to explore nodes in the tree, have been implemented. Some interactive tools for tree-based survival analysis were described by LeBlanc (1990).

APPENDIX: PROOF OF THEOREM 5.1

We use the notation as defined in Section 5 and earlier. Let

$$\bar{y}_N(v) = \frac{1}{\eta_N(v)} \sum_{\mathbf{X}_i \in v} Y_i$$

and

$$\bar{\mu}_N(v) = \frac{1}{\eta_N(v)} \sum_{\mathbf{X}_i \in v} E[Y_i | X_i].$$

First, we quote an important lemma.

Lemma. Assume that $\lim_{N\to\infty} k_N = \infty$ and Y is bounded; then for every polyhedron B with a uniformly bounded number of faces in $\mathbf{R}^M$ and every $\varepsilon > 0$ and $c > 0$,

$$\lim_N N^c P\Bigg(|\bar{y}_N(v) - \bar{\mu}_N(v)| > \varepsilon \text{ for some } v \in \mathcal{B} \\ \text{such that } v \subset B \text{ and } p_N(v) \geq k_N \frac{\log N}{N} \Bigg) = 0.$$

This is Lemma 12.27 from Breiman et al. (1984), restricted to the special case of bounded random variables.

Proof of Theorem 5.1. $\delta I_{\{T\leq t\}}$ and $I_{\{T\leq t\}}$ are bounded. Therefore, we can obtain the previous lemma result pointwise in t for $\varepsilon = \varepsilon'/2$. Because $H_N(t, v)$, $H_{N1x}(t, v)$, $\bar{H}_N(t, v)$, and $\bar{H}_{N1}(t, v)$ are bounded and monotone, the pointwise result can be improved to a uniform result on $[0, \tau]$. Thus

$$\lim_N N^c P\Bigg(\sup_{t\leq\tau} |H_N(t, v) - \bar{H}_N(t, v)| > \varepsilon' \text{ for some } v \in \mathcal{B} \\ \text{such that } v \subset B \text{ and } p_N(v) \geq k_N \frac{\log N}{N} \Bigg) = 0,$$

and

$$\lim_N N^c P\Bigg(\sup_{t\leq\tau} |H_{1N}(t, v) - \bar{H}_{1N}(t, v)| > \varepsilon' \text{ for some } v \in \mathcal{B} \\ \text{such that } v \subset B \text{ and } p_N(v) \geq k_N \frac{\log N}{N} \Bigg) = 0.$$

Also

$$\begin{aligned} &P(|\hat{\Lambda}_N(t, v) - \bar{\Lambda}_N(t, v)| > \varepsilon/2) \\ &\quad = P\left(\left| \int_0^t \frac{dH_{1N}(s, v)}{1 - H_N(s^-, v)} - \int_0^t \frac{d\bar{H}_{1N}(s, v)}{1 - \bar{H}_N(s^-, v)} \right| > \varepsilon/2 \right) \\ &\quad \leq P\left(\left| \int_0^t \frac{1}{1 - H_N(s^-, v)} [dH_{1N}(s, v) - d\bar{H}_{1N}(s, v)] \right| > \varepsilon/4 \right) \end{aligned}$$

$$
\begin{aligned}
&+ P\left(\left| \int_0^t \left(\frac{1}{1 - H_N(s^-, v)} \right.\right.\right. \\
&\qquad \left.\left.\left. - \frac{1}{1 - \bar{H}_N(s^-, v)} \right) d\bar{H}_{1N}(s, v) \right| > \varepsilon/4 \right) \\
&\le P(|\sup_{t \le \tau} |H_{1N}(t, v) - \bar{H}_{1N}(t, v)| > \varepsilon/4M_1) \\
&\quad + P(|\sup_{t \le \tau} |H_N(t^-, v) - \bar{H}_N(t^-, v)| > \varepsilon/4M_2),
\end{aligned}
$$

for M_1, M_2 and N sufficiently large by integration by parts. Therefore, pointwise

$$
\lim_N N^c P\left(|\hat{\Lambda}_N(t, v) - \bar{\Lambda}_N(t, v)| > \varepsilon/2 \text{ for some } v \in \mathcal{B} \right.
$$
$$
\left. \text{such that } v \subset B \text{ and } p_N(v) \ge k_N \frac{\log N}{N} \right) = 0.
$$

The uniform result follows because $\hat{\Lambda}(t, v)$ and $\bar{\Lambda}(t, v)$ are bounded and monotone.

[*Received October 1990. Revised August 1992.*]

REFERENCES

Akaike, H. (1974), "A New Look at Model Identification," *IEEE Transactions on Automatic Control,* 19, 716–723.

Albain, K., Crowley, J., LeBlanc, M., and Livingston, R. (1990), "Determinants of Improved Outcome in Small-Cell Lung Cancer: An Analysis of the 2,580-Patient Southwest Oncology Group Data Base," *Journal of Clinical Oncology,* 8, 1563–1574.

Becker, R., Chambers, J., and Wilks, A. (1988), *The New S Language,* Pacific Grove, CA: Wadsworth International Group.

Bloch, D. A., and Segal, M. R. (1989), "Empirical Comparison of Approaches to Forming Strata," *Journal of the American Statistical Association,* 84, 897–905.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984), *Classification and Regression Trees,* Belmont, CA: Wadsworth International Group.

Breslow, N., and Crowley, J. (1974), "A Large Sample Study of the Life Table and Product Limit Estimates Under Random Censorship," *The Annals of Statistics,* 2, 437–453.

Butler, J., Gilpin, E., Gordon, L., and Olshen, R. (1989), "Tree-Structured Survival Analysis, II," technical report. Stanford University, *Dept. of Biostatistics.*

Ciampi, A., Hogg, S., McKinney, S., and Thiffault, J. (1988), "RECPAM: A Computer Program for Recursive Partition and Amalgamation for Censored Survival Data," *Computer Methods and Programs in Biomedicine,* 26, 239–256.

Ciampi, A., Thiffault, J., Nakache, J-P., and Asselain, B. (1986), "Stratification by Stepwise Regression, Correspondence Analysis and Recursive Partition," *Computational Statistics and Data Analysis,* 4, 185–204.

Clark, L., and Pregibon, D. (1991), "Tree-Based Models," in *Statistical Models in S,* eds. J. M. Chambers and T. Hastie, Pacific Grove, CA: Wadsworth International Group.

Cox, D. R. (1972), "Regression Models and Life Tables," *Journal of the Royal Statistical Society,* Ser. B, 34, 187–200.

Crawford, S. L. (1987), "Resampling Strategies for Recursive Partitioning Classification with the CART Algorithm," Ph.D. Dissertation, Stanford University, Dept. of Statistics.

——— (1989), "Extensions to the CART Algorithm," *International Journal of Man-Machine Studies,* 31, 197–217.

Davis, R., and Anderson, J. (1989), "Exponential Survival Trees," *Statistics in Medicine,* 8, 947–962.

Efron, B. (1983), "Estimating the Error Rate of a Prediction Rule: Improvements on Cross-Validation," *Journal of the American Statistical Association,* 78, 316–331.

Friedman, J. H. (1991), "Multivariate Adaptive Regression Splines" (with discussion), *The Annals of Statistics,* 19, 1–141.

Fries, D. A., Bloch, D. A., Segal, M. R., Spitz, P. W., Williams, C., and Lane, N. E. (1988), "Postmarketing Surveillance in Rheumatology: Analysis of Purpura and Upper Abdominal Pain," *The Journal of Rheumatology,* 15, 348–355.

Gill, R. (1980), *Censoring and Stochastic Integrals,* Mathematical Centre Tracts, 124, Mathematisch Centrum, Amsterdam.

Gordon, L., and Olshen, R. (1980), "Consistent Nonparametric Regression From Recursive Partitioning Schemes," *Journal of Multivariate Analysis,* 10, 611–627.

——— (1984), "Almost Surely Consistent Nonparametric Regresssion From Recursive Partitioning Schemes," *Journal of Multivariate Analysis,* 15, 147–163.

——— (1985), "Tree-Structured Survival Analysis," *Cancer Treatment Reports,* 69, 1065–1069.

Halpern, J. (1982), "Maximally Selected Chi-Squared Statistics for Small Samples," *Biometrics,* 38, 1017–1023.

Harrington, D., and Fleming, T. (1982), "A Class of Rank Test Procedures for Censored Survival Data," *Biometrika,* 69, 553–566.

Jespersen, N. C. B. (1986), "Dichotomizing a Continuous Covariate in the Cox Regression Model," Technical Report, University of Copenhagen, Statistical Research Unit and Institute of Mathematical Statistics.

Kwak, L. W., Halpern, J., Olshen, R. A., and Horning, S. J. (1990), "Prognostic Significance of Actual Dose Intensity in Diffuse Large-Cell Lymphoma: Results of a Tree-Structured Survival Analysis," *Journal of Clinical Oncology,* 8, 963–977.

LeBlanc, M. (1989), "Regression Trees for Censored Survival Data," unpublished Ph.D. dissertation, University of Washington, Dept. of Biostatistics.

——— (1990), "Tree-Based Tools for Survival Data," in *Proceedings of the XV International Biometrics Conference,* pp. 123–133.

LeBlanc, M., and Crowley, J. (1992), "Relative Risk Trees for Censored Survival Data," *Biometrics,* 48, 411–425.

Miller, R., and Siegmund, D. (1982), "Maximally Selected Chi-Squared Statistics," *Biometrics,* 38, 1011–1016.

Morgan, J., and Sonquist, J. (1963), "Problems in the Analysis of Survey Data and a Proposal," *Journal of the American Statistical Association,* 58, 415–434.

Nelson, W. (1969), "On Estimating the Distribution of Random Vectors When Only the Coordinate is Observable," *Technometrics,* 12, 923–924.

Owen, A. (1991), Discussion of "Multivariate Adaptive Regression Splines," by J. H. Friedman, *The Annals of Statistics,* 19, 102–112.

Segal, M. (1988), "Regression Trees for Censored Data," *Biometrics,* 44, 35–48.

Shorack, G., and Wellner, J. (1986), *Empirical Processes and Applications to Statistics,* New York: John Wiley.

Stone, M. (1974), "Choice and Assessment of Statistical Predictions," *Journal of the Royal Statistical Society,* Ser. B, 36, 111–133.

Therneau, T., Grambsch, P., and Fleming, T. (1990), "Martingale-Based Residuals for Survival Models," *Biometrika,* 77, 147–160.

Tsiatis, A. (1975), "A Nonidentifiability Aspect of the Problem of Competing Risks," *Proceedings of the National Academy of Sciences of the United States of America,* 74, 20–22.