BIOMETRICS 48, 411–425
June 1992

Relative Risk Trees for Censored Survival Data

Michael LeBlanc

Department of Preventive Medicine and Biostatistics, University of Toronto,
Toronto M5S 1A8, Canada

and

John Crowley

Fred Hutchinson Cancer Research Center, 1124 Columbia Street,
Seattle, Washington 98104, U.S.A.

SUMMARY

A method is developed for obtaining tree-structured relative risk estimates for censored survival data. The first step of a full likelihood estimation procedure is used in a recursive partitioning algorithm that adopts most aspects of the widely used Classification and Regression Tree (CART) algorithm of Breiman et al. (1984, *Classification and Regression Trees*, Belmont, California: Wadsworth). The performance of the technique is investigated through simulation and compared to the tree-structured survival methods proposed by Davis and Anderson (1989, *Statistics in Medicine* **8**, 947–961) and Therneau, Grambsch, and Fleming (1990, *Biometrika* **77**, 147–160).

1. Introduction

Tree-based methods for regression, and especially classification, are becoming popular alternatives to linear regression and linear discriminant analysis. Trees generally require fewer assumptions than classical methods and handle a wide variety of data structures. They provide another way of understanding the predictive structure of the data for both statistically and nonstatistically oriented researchers. These methods (often called recursive partitioning) were originally developed by Morgan and Sonquist (1963); the Classification and Regression Tree (CART) algorithm and monograph of Breiman et al. (1984) greatly advanced the technology, and stimulated wide interest in tree-based techniques.

Tree-based methods adaptively partition the covariate space into regions and the data into groups. For each ordered covariate and split of the form "$X_j \leq c$ or $X_j > c$" some measure of separation in the response distribution (for instance, a likelihood ratio test statistic) between the two groups is calculated. More general splits are possible for non-ordered categorical variables. The covariate and the split point that best separates the groups are chosen and this same procedure is applied recursively to the resulting groups until many disjoint regions, each containing only a few observations, have been formed. The resulting model can be represented as a binary tree. After a large tree is grown, there are rules for recombining nodes and for choosing the size of the tree. While we review aspects of tree-based methods necessary to describe our procedure, for an extensive discussion see Breiman et al. (1984).

Tree-based methods could be a useful alternative to the classical linear proportional

Key words: Censored data; Proportional hazards; Regression trees.

hazards model of Cox (1972) for the exploration of survival data. The partitioning of the covariate space creates "bins" of observations that are assumed to be approximately homogeneous. This permits the use of one-sample tools for censored survival, such as the Kaplan–Meier estimator or other simple functionals such as quantiles, to compare prognosis between the "bins" represented by nodes in the tree. Also, the combination of binning and the interpretability of the tree-structured description make recursive partitioning well suited for developing prognostic stratifications that are used in the design of clinical trials. Several tree-based tools have been proposed for censored survival data (Gordon and Olshen, 1985; Ciampi et al., 1986; Segal, 1988; Davis and Anderson, 1989).

Our tree-structured method adopts the proportional hazards model which specifies the following hazard function at time t, for an individual with covariate vector $\mathbf{x}$:

$$\lambda(t \mid \mathbf{x}) = \lambda_0(t) s(\mathbf{x}),$$

where $s(\mathbf{x}) \geqslant 0$ and $\lambda_0(t)$ is the baseline hazard. Traditionally $s(\mathbf{x})$ is a log-linear function of a vector of parameters, but recently the model has been extended to include nonparametric covariate effects. The linear function is typically replaced by an additive function (Tibshirani and Hastie, 1987; O'Sullivan, 1988; Gentleman and Crowley, 1991). We will present a method for obtaining tree structures that represent the relative risk function, $s(\mathbf{x})$.

Since recursive partitioning involves evaluation of a large number of splits, iterative computation at each split point would typically not be computationally feasible, especially if tree-based modeling is to be carried out in an interactive data analysis environment as proposed by Becker, Clark, and Pregibon (1989) and LeBlanc (1990). Therefore, the method grows and prunes a tree using only the first step of a full likelihood estimation procedure for the proportional hazards model. After a tree is chosen, full likelihood estimates are obtained by iteration.

We adopt much of the CART "engineering" of Breiman et al. (1984), the current standard for recursive partitioning algorithms. This methodology includes the cost-complexity pruning algorithm, which efficiently yields trees that perform best in terms of residual error (deviance for our model) for their size. In addition, the tree-growing process is cross-validated to estimate prediction error for a sequence of models. The tree model that minimizes or comes close to minimizing estimated prediction error is chosen. In these respects our algorithm is similar to some other proposals that extend the CART algorithm to right-censored data. Gordon and Olshen (1985) use distances between estimated distribution functions based on L_p and L_p Wasserstein metrics, and Davis and Anderson (1989) use exponential log-likelihood to define cost for a node. However, while the one-step full likelihood method is similar to these other two methods, it has the advantage of being based on the popular proportional hazards model. It should be more generally applicable than the method based on parametric exponential likelihood, and it is as easy to implement and as computationally efficient. Other tree-based methods for survival data proposed by Segal (1988) and Ciampi et al. (1986) are also interpretable but are different in spirit since they are based solely on between-node separation based on two-sample log-rank test statistics rather than some measure of within-node error such as deviance; since the methods do not use a within-node measure of error, they propose alternatives to the CART pruning algorithm and to cross-validation to select tree size.

We present the results of a simulation study that investigates the performance of the full likelihood method and compares it to the performance of the exponential likelihood-based method of Davis and Anderson (1989) and a proposal of Therneau, Grambsch, and Fleming (1990), who use martingale-type "residuals" directly in the CART regression algorithm. Finally, an example is given based on data from a randomized clinical trial.

2. The Likelihood

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 $\mathbf{X} = (X_1, X_2, \ldots, X_M)$ is a vector of M covariates. Suppose U is the true survival time having cumulative distribution function F, and V is the true censoring time with cumulative distribution function G. Let $\delta = I_{\{U \leq V\}}$, where $I_{\{\cdot\}}$ is the indicator function of the set $\{\cdot\}$, and the observed time, $T = \min(U, V)$. Assume also that the U and V are independent given $\mathbf{X}$. The learning sample consists of the set of independent, identically distributed vectors $\{(t_i, \delta_i, \mathbf{x}_i)\colon i = 1, 2, \ldots, N\}$.

Typically inference for the proportional hazards model is based on the partial likelihood. However, if the baseline cumulative hazard is known, estimation and model selection based on the full likelihood are desirable. The full likelihood of the learning sample for a tree T can be expressed as

$$L = \prod_{h \in \tilde{T}} \prod_{i \in S_h} \lambda_h(t_i)^{\delta_i} e^{-\Lambda_h(t_i)},$$

where $\tilde{T}$ is the set of terminal nodes; S_h is the set of observation labels, $\{i\colon \mathbf{x}_i \in \chi_h\}$, for observations in the region χ_h corresponding to node h; (t_i, δ_i) is the vector of observation time and failure indicator for individual i; and $\lambda_h(t)$ and $\Lambda_h(t)$ are the hazard and cumulative hazard functions for node h. Assume that the proportional hazards model

$$\lambda_h(t) = \theta_h \lambda_0(t)$$

is true, where θ_h is a nonnegative parameter and $\lambda_0(t)$ is the baseline hazard. It follows that the likelihood for the data given tree T is

$$L = \prod_{h \in \tilde{T}} \prod_{i \in S_h} (\lambda_0(t_i)\theta_h)^{\delta_i} e^{-\Lambda_0(t_i)\theta_h},$$

where $\Lambda_0(t)$ is the baseline cumulative hazard function. Given the baseline cumulative hazard, the maximum likelihood estimates of $\{\theta_h\colon h \in \tilde{T}\}$ are

$$\tilde{\theta}_h = \frac{\sum_{i \in S_h} \delta_i}{\sum_{i \in S_h} \Lambda_0(t_i)}.$$

In practice, the cumulative hazard is not known. However, a natural estimator of the cumulative hazard given estimates $\hat{\theta}_h$,

$$\hat{\Lambda}_0(t) = \sum_{i: t_i \leq t} \frac{\delta_i}{\sum_{h \in \tilde{T}} \sum_{i: t_i \geq t, i \in S_h} \hat{\theta}_h},$$

is due to Breslow (1972). It can be shown that if $\Lambda(t)$ is replaced by $\hat{\Lambda}_0(t)$ in the full likelihood score equations, one obtains the partial likelihood score equations. An alternating estimation procedure can be used to estimate $\Lambda_0(t)$ and the $\{\theta_h\colon h \in \tilde{T}\}$. First the Breslow cumulative hazard estimate for iteration j,

$$\hat{\Lambda}_0^j(t) = \sum_{i: t_i \leq t} \frac{\delta_i}{\sum_{h \in \tilde{t}} \sum_{i: t_i \geq t, i \in S_h} \hat{\theta}_h^j}, \tag{1}$$

is calculated using the current estimates, $\hat{\theta}_h^j$, of θ_h. Next, the estimate $\hat{\theta}_h^{j+1}$ of θ_h,

$$\hat{\theta}_h^{j+1} = \frac{\sum_{i \in S_h} \delta_i}{\sum_{i \in S_h} \hat{\Lambda}_0^j(t_i)}, \tag{2}$$

is calculated using the current estimate $\hat{\Lambda}_0^j(t_i)$. The two steps are repeated until convergence. At convergence, the estimates of $\{\theta_h: h \in \tilde{T}\}$ are not uniquely defined; only the ratios of the estimates between nodes are unique. This technique was used by Clayton (1983) and Clayton and Cuzick (1985) to fit linear proportional hazards models and by Gentleman and Crowley (1991) to fit additive proportional hazards models.

Only the first iteration will be used in the recursive partitioning procedure to grow the tree and select the tree size. The Breslow estimator evaluated at $\{\hat{\theta}_h^j = 1: h \in \tilde{T}\}$, which is the Nelson (1969) cumulative hazard estimator, is used. The one-step estimate of θ_h is

$$\hat{\theta}_h^1 = \frac{\sum_{i \in S_h} \delta_i}{\sum_{i \in S_h} \hat{\Lambda}_0^1(t_i)},$$

which can be interpreted as the observed number of deaths divided by the expected number of deaths in node h under the assumption of no structure in survival times. Hence, even the one-step procedure gives interpretable quantities for each node.

The full likelihood deviance measures how well the tree fits the data. The deviance for node h is

$$R(h) = 2\{L_h(\text{saturated}) - L_h(\tilde{\theta}_h)\},$$

where L_h(saturated) is the log-likelihood for the saturated model that allows one parameter for each observation, and $L_h(\tilde{\theta}_h)$ is the maximized log-likelihood when $\Lambda_0(t)$ is known. It can be shown that the deviance residual for an observation i in node h is

$$d_i = 2\left[\delta_i \log\left(\frac{\delta_i}{\Lambda_0(t_i)\hat{\theta}_h}\right) - (\delta_i - \Lambda_0(t_i)\hat{\theta}_h)\right].$$

The residual is equivalent to the deviance residual based on the Poisson model with response δ_i and mean $\tilde{\mu}_i = \Lambda_0(t_i)\tilde{\theta}_h$. The recursive partitioning procedure calculates one-step deviance residuals with the Nelson estimate substituted for $\Lambda_0(t)$, and with the one-step parameter estimates $\{\hat{\theta}_h^1: h \in \tilde{T}\}$. The connection between the proportional hazards full likelihood and the Poisson model likelihood has been used by several authors including Clayton (1983) and Clayton and Cuzick (1985).

3. The Algorithm

The algorithm adopts the important aspects of the CART algorithm. A large tree is grown to avoid missing important structure. The cost-complexity pruning algorithm of CART obtains an optimal sequence of subtrees (subtrees are obtained by removing branches from a tree). Finally, an estimate of expected one-step deviance is calculated for each of the pruned subtrees by cross-validation and the tree that minimizes the estimated deviance is chosen as the best tree.

Recursive partitioning algorithms split the covariate space based on a rule that maximizes some measure of improvement. Our algorithm will split the data and the covariate space into regions that maximize the reduction in one-step deviance realized by the split. Many types of partitions could be considered; however, we will consider only splits on a single variable. All possible splits for each of the covariates are evaluated and the variable and split point resulting in the greatest reduction in the one-step deviance are chosen. Usually, there is a rule regarding the minimum size of a node, since if very small nodes are permitted the algorithm often splinters off small groups of observations resulting in trees that do not validate well. Let N be the total number of observations in the learning sample. The improvement for split s at node h into daughter nodes $l(h)$ and $r(h)$ is

$$R(s, h) = R(h) - [R(l(h)) + R(r(h))],$$

where

$$R(h) = \frac{1}{N} \sum_{i \in S_h} \left[\delta_i \log \left(\frac{\delta_i}{\hat{\Lambda}_0^1(t_i)\hat{\theta}_h} \right) - (\delta_i - \hat{\Lambda}_0^1(t_i)\hat{\theta}_h) \right].$$

The binary splitting continues until a large binary tree is grown and there are only a few observations in each node.

In a simulation experiment it has been shown that the performance of the reduction in the one-step deviance is very similar to the log-rank test statistic used in the tree-based methods of Segal (1988) and Ciampi et al. (1986) (LeBlanc, unpublished Ph.D. thesis, Department of Biostatistics, University of Washington, 1989). Our splitting statistic is chosen because of the direct analogy with splitting based on the reduction of mean residual sums of squares in the CART regression algorithm.

3.1 *Pruning and Tree Selection*

The cost-complexity of a tree is defined (Breiman et al., 1984) to be

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

for a nonnegative complexity parameter α, where $R(h)$ is the impurity of node h defined above.

The cost-complexity measure controls the tradeoff between the size or complexity of the tree, and how well the tree fits the data. If the complexity parameter α is large the tree that minimizes the cost complexity is small and as α decreases the tree that minimizes the cost complexity increases in size. We will choose trees that minimize the cost complexity measure just as is done in CART; these are called optimally pruned subtrees. In the next definition, the symbol "$\preceq$" means "is a subtree of."

Definition 3.1. T_1 is an optimally pruned subtree of T for complexity parameter α, if

$$R_\alpha(T_1) = \min_{T' \preceq T} R_\alpha(T'),$$

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

Breiman et al. (1984) show that for the cost-complexity measure there is a unique smallest optimally pruned subtree for any complexity parameter α. They also show that as α increases, the optimal sequence of subtrees is a nested sequence of trees and that there is an efficient algorithm for obtaining the optimal sequence of subtrees.

The expected one-step deviance for the pruned trees is estimated by V-fold cross-validation. The data $\mathscr{L}$ are divided up into V sets $\mathscr{L}_v$ and subsamples $\mathscr{L}_{(v)} = \mathscr{L} - \mathscr{L}_v$ of about equal size, and trees T_v are grown from the subsamples $\mathscr{L}_{(v)}$. For any α, an optimally pruned subtree, $T_v(\alpha)$, and estimates $\hat{\theta}_h^1(v)$: $h \in \tilde{T}_v(\alpha)$ are obtained. For each tree T_v approximately $1/V$ of the data is run down the tree. We evaluate the performance of the tree-based model generated with the sample $\mathscr{L}_{(v)}$ with the sample $\mathscr{L}_v$. The cross-validated deviance residual for individual i not in the sample used to grow the tree is

$$d_i(\delta_i, \hat{\theta}_h^1(v)) = 2 \left[\delta_i \log \left(\frac{\delta_i}{\hat{\Lambda}_0^1(t_i)\hat{\theta}_h^1(v)} \right) - (\delta_i - \hat{\Lambda}_0^1(t_i)\hat{\theta}_h^1(v)) \right],$$

where $\hat{\Lambda}_0^1(t_i)$ is based on $\mathscr{L}$. Let α^* be the value of complexity parameter α that minimizes the average cross-validated deviance residuals for trees $T_v(\alpha)$ over the V subsamples. Note that in the cross-validation procedure we have described, the entire tree-growing process is repeated for each subsample. If the tree-structured model was fixed so that only the estimates for each node were recalculated, then the procedure would underestimate the expected deviance of the trees. For a detailed description of cost-complexity pruning and selection of tree size, readers are again referred to Breiman et al. (1984).

A problem arises when calculating the cross-validated estimates for censored data. If all the observations in node h are censored in the subsample used to grow the tree, the estimate $\hat{\theta}_h$ is 0. Now if some of the observations in the validation sample for that node are uncensored, the cross-validation estimate of the expected deviance is infinite. Since zero hazard estimates and infinite deviance are unrealistic, an adjustment is needed. In the parametric setting T. Therneau (personal communication) has suggested shrinkage estimators to avoid the problem of estimated hazard functions of zero. However, a simple ad hoc solution is to replace nodes with zero observed deaths with .5, similar to what is suggested by Davis and Anderson (1989) for exponential likelihood. Then, the estimate of θ_k for cross-validation is

$$\hat{\theta}_h = \frac{1}{2\sum_{i \in S_h} \hat{\Lambda}_0^1(t_i)}$$

for a node h that has no observed deaths.

After choosing a tree $T(\alpha^*)$ minimizing the cross-validated estimate of the expected one-step deviance, maximum likelihood estimates of the relative risk between nodes are obtained by iterating on equations (1) and (2), which does not involve any matrix inversion. Convergence is rapid; this is unlike the additive model where convergence of this method can be extremely slow (Gentleman and Crowley, 1991).

It is well known that the size of the trees selected by minimizing the cross-validation estimate of prediction error can be quite variable. The addition of some technique to choose simpler trees that perform not substantially worse in terms of prediction error than the tree minimizing the cross-validation estimate of prediction error, such as the "1 SE" rule of Breiman et al. (1984), would be useful. However, further study of such rules is needed, especially in the censored data setting. One likelihood-based approach is suggested by Davis and Anderson (1989).

4. One-Step Scores

One-step weighted least squares scores are an alternative to the full likelihood deviance measure of tree performance. The adjusted dependent variable (McCullagh and Nelder, 1983) and weights can be calculated if $\Lambda_0(t)$ is assumed to be known. The adjusted dependent variable, y_i, for individual i in node h in our parameterization is

$$y_i = \frac{\delta_i}{\Lambda_0(t_i)},$$

and the weight is

$$w_i = \frac{\Lambda_0(t_i)}{\theta_h}.$$

The impurity of the node is based on the weighted sum of squares for the node. Again, substitute the Nelson estimate of the cumulative hazard $\hat{\Lambda}_0^1(t)$ for $\Lambda_0(t)$. Let w_{0i} be the weight function above defined at $\theta_h = 1$. Then the weighted least squares score for node h is

$$R(h) = \frac{1}{N} \sum_{i \in S_h} w_{0i}(y_i - \theta_h)^2$$

$$= \frac{1}{N} \sum_{i \in S_h} \frac{(\delta_i - \hat{\Lambda}_0^1(t_i)\theta_h)^2}{\hat{\Lambda}_0^1(t_i)}. \tag{3}$$

The factor $\delta_i - \hat{\Lambda}_0^1(t_i)\theta_h$ is called a martingale residual by Therneau et al. (1990). It can be interpreted as the number of observed deaths for individual i minus an estimate of the expected number of deaths under the assumption of the tree-structured proportional hazards model.

The value of θ_h that minimizes (3) is

$$\hat{\theta}_h = \frac{\sum_{i \in S_h} \delta_i}{\sum_{i \in S_h} \hat{\Lambda}_0^1(t_i)},$$

which is the maximum likelihood estimator of θ_h with $\hat{\Lambda}_0^1(t)$ substituted for $\Lambda_0(t)$.

The weighted least squares approach permits easy implementation, since the CART regression algorithm need only be changed to include a weight function for each observation. In addition, there are simple updating formulas for weighted least squares for rapidly calculating splitting statistics.

Therneau et al. (1990) use martingale residuals directly in the CART regression algorithm with squared error loss. The mean of the martingale residuals in a node is the summary statistic for their proposal. However, if the proportional hazards assumption is roughly correct, the one-step estimates of θ_h developed here seem to have a better interpretation since each is the ratio of observed to expected deaths in the node under the assumption of no structure.

4.1 *Alternative Weights*

Different weights can be easily incorporated into the least squares scores to allow possibly more robust tree-structured survival analysis than that based on the proportional hazards model.

Let $w(t)$ be a weight function at time t. An alternative least squares score for node h is

$$R(h) = \frac{1}{N} \sum_{i \in S_h} \frac{(w(t_i)\delta_i - \theta_h \int_0^\infty w(s)Y_i(s)\, d\hat{\Lambda}_0^1(s))^2}{\int_0^\infty w(s)Y_i(s)\, d\hat{\Lambda}_0^1(s)},$$

where $Y_i = I_{\{T_i \geq s\}}$. The estimate of θ_h that minimizes $R(h)$ is

$$\hat{\theta}_{h,w}^1 = \frac{\sum_{i \in S_h} w(t_i)\delta_i}{\sum_{i \in S_h} \int_0^\infty w(s)Y_i(s)\, d\hat{\Lambda}_0^1(s)}. \tag{4}$$

The estimate $\hat{\theta}_{h,w}^1$ is an estimate of the ratio of the weighted number of observed deaths in node h to the weighted number of expected deaths under the assumption of no structure. The weight function can be chosen to focus on survival differences of interest between nodes. For instance, if one wanted estimation insensitive to late differences, a weight function that is a decreasing function of time would be appropriate. Two examples of weight functions that may be useful are $w(t) = \bar{F}_{\text{est}}(t)$, where $\bar{F}_{\text{est}}(t)$ is the Kaplan–Meier estimate of the survival function for the entire learning sample, and $w(t) = I_{\{t \leq c\}}$, where c is some positive constant. The resulting estimates $\hat{\theta}_{h,w}^1$ have a connection with the estimates studied by Sasieni (unpublished Ph.D. thesis, University of Washington, 1989).

5. Simulation Experiment

A simulation study was done to investigate the performance of the one-step full likelihood method, and to compare it to the martingale residual technique proposed by Therneau et al. (1990) and the exponential log-likelihood method of Davis and Anderson (1989). Since the recursive partitioning techniques do not use the same loss function, an estimate of expected deviance based on the true survival distribution was used to assess the performance of the procedures.

5.1 *Method*

We simulated data from a total of five models. In each case there were five covariates, $X_1, \ldots, X_5$, independent and uniformly distributed over the unit interval. Members of the $H^\rho(\psi, \rho)$ family (Harrington and Fleming, 1982),

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

were used to generate survival times, and censoring times were chosen to be uniformly distributed on $(0, \gamma)$.

We considered five different survival models where $\theta = \log(\psi)$:

$$\begin{array}{lll} \textbf{A:} & \theta_i = 0, & \rho = 0; \\ \textbf{B:} & \theta_i = I_{\{x_{1i} \leq .5 \cap x_{2i} > .5\}}, & \rho = 0; \\ \textbf{C:} & \theta_i = 3.0x_{1i} + 1.0x_{2i}, & \rho = 0; \\ \textbf{D:} & \theta_i = I_{\{x_{1i} \leq .5 \cap x_{2i} > .5\}} + .367, & \rho = 1; \\ \textbf{E:} & \theta_i = 3.0x_{1i} + 1.0x_{2i} + .367, & \rho = 1. \end{array}$$

For models B, C, D, and E, the survival distribution depends on two of the five measurement variables. Models B and D specify one region of the measurement space to correspond to poor prognosis. In models C and E, θ is linear in x_1 and x_2; for models A, B, and C, the survival distributions are exponential and hence they are proportional hazards models. Models D and E have decreasing hazard ratios over time between points in the measurement space. Parameters for models D and E were chosen such that at any point in the covariate space, the median survival is the same as for models B and C, respectively.

The five models were examined for a sample size of $n = 250$ observations. Only model C was simulated with a sample size of $n = 500$ observations. Two hundred fifty samples were generated from each survival model and censoring distribution. The minimum node size permitted for splitting was 20 observations.

Each of the methods selects the pruned subtree minimizing the tenfold cross-validation estimate of the prediction error for that method.

The estimates of the expected losses for the trees were calculated by sending 2,500 observations generated by the same model down the pruned trees.

5.2 *Results*

Results on the number of terminal nodes for the three methods are presented in Table 1, and on the estimated expected deviances in Table 2.

No structure If the failure time distribution does not depend on the measurement variables, the recursive partitioning technique should not falsely detect structure. For data generated

Table 1

The number of terminal nodes for one-step full likelihood (FL), *martingale residuals* (MR), *and exponential likelihood* (EX) *methods for data generated from models* A–E

Model and method	No censoring						50% censoring					
	Number of terminal nodes						Number of terminal nodes					
	1	2	3	4	5	≥6	1	2	3	4	5	≥6
A ($n = 250$)												
FL	94.4	3.2	1.2	.8	.4	.0	95.6	3.6	.4	.4	.0	.0
MR	94.8	1.6	1.6	1.6	.4	.0	93.6	4.0	1.2	1.2	.0	.0
EX	93.2	4.0	1.2	1.2	.4	.0	94.8	3.2	1.6	.0	.0	.4
B ($n = 250$)												
FL	10.8	4.4	64.0	11.6	4.4	4.8	33.6	10.4	38.4	10.8	4.0	2.8
MR	26.4	16.4	28.0	14.0	7.2	8.0	29.2	14.0	38.8	9.2	5.2	3.6
EX	12.8	4.4	60.8	13.6	3.6	4.8	35.2	10.4	37.6	11.2	4.0	1.6
C ($n = 250$)												
FL	.0	6.0	18.4	9.2	14.4	52.0	.0	34.0	26.4	12.0	11.6	16.0
MR	.0	17.6	8.8	12.0	14.4	47.2	.0	36.8	23.2	13.2	8.4	18.4
EX	.0	10.0	18.0	11.6	15.2	45.2	.0	33.2	26.0	13.6	6.8	20.4
D ($n = 250$)												
FL	79.6	7.2	8.0	2.8	1.2	1.2	70.4	9.6	10.4	6.0	1.6	2.0
MR	78.8	8.8	6.8	2.0	1.6	2.0	68.0	10.4	10.0	5.6	2.8	3.2
EX	74.4	6.8	4.4	3.2	4.8	6.4	74.8	8.8	8.0	3.6	2.4	2.4
E ($n = 250$)												
FL	2.0	55.6	15.2	10.0	6.8	10.4	2.4	61.6	15.2	6.8	5.6	8.4
MR	1.2	58.8	12.8	9.6	6.0	11.6	2.0	53.2	14.8	12.0	6.4	11.6
EX	38.8	25.2	12.0	7.2	6.0	10.8	5.6	60.0	12.0	7.2	6.8	8.4

from model A, Table 1 shows that for all three techniques more than 93% of the trees consisted only of the root node.

Model A was also investigated with uneven censoring. For $x_{1i} \leqslant .5$ the censoring distribution was $U(0, \gamma_1)$ and for $x_{1i} > .5$ 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, respectively. Table 2 shows that there were no important differences between expected deviance for the methods for uneven censoring compared to no censoring. Also, there were no significant differences in the distribution of tree complexities. The more general problem of censoring related to covariates was not investigated.

Structure The simplest tree representing the structure for model B would have three terminal nodes with splits at about $x_1 = .5$ and $x_2 = .5$. Although such a parsimonious tree may not be found, Table 1 shows that for the three methods and for both the uncensored and 50% censoring cases, the modal number of terminal nodes is three. The martingale residual (MR) technique selected the largest proportion of trees with no structure, 26% in the uncensored case. Table 2 shows that for uncensored data the estimated expected deviance is larger for the MR technique than for the other techniques for uncensored data.

In the case of 50% censoring, differences in expected deviances between the MR, full likelihood (FL), and exponential likelihood (EX) methods are not evident based on the number of simulations performed.

For uncensored data generated from model C, Table 1 shows that the tree size is extremely variable for the three methods. With 50% censoring the methods again have similar distributions of tree sizes with a smaller average selected tree size and increased expected deviance.

Table 2
Estimated expected deviances for one-step full likelihood (FL), *martingale residuals* (MR), *and exponential likelihood* (EX) *methods for data generated from models* A–E. *Standard errors of estimates are given in parentheses.*

	Recursive partitioning technique		
Model	FL	MR	EX
A ($n = 250$)			
No censoring	292.1 (.79)	291.4 (.69)	292.0 (.79)
50% censoring	295.0 (1.04)	295.3 (1.10)	295.3 (1.17)
Uneven censoring	292.2 (.80)	292.0 (.78)	292.7 (.89)
	Expected true model deviance = 288.6		
B ($n = 250$)			
No censoring	313.7 (1.42)	321.7 (1.07)	314.1 (1.44)
50% censoring	331.2 (1.86)	331.0 (1.74)	331.1 (1.70)
	Expected correct model deviance = 288.6		
C ($n = 250$)			
No censoring	345.3 (1.24)	349.3 (1.34)	344.2 (1.21)
50% censoring	383.9 (2.18)	374.3 (1.97)	380.3 (2.13)
	Expected correct model deviance = 288.6		
D ($n = 250$)			
No censoring	325.1 (.67)	326.1 (.69)	327.8 (.72)
50% censoring	327.6 (.77)	328.1 (.81)	327.0 (.74)
	Expected correct model deviance = 308.2		
E ($n = 250$)			
No censoring	342.7 (.89)	342.7 (.82)	357.4 (1.40)
50% censoring	347.1 (1.07)	347.1 (1.05)	347.6 (1.08)
	Expected correct model deviance = 308.2		
C ($n = 500$)			
No censoring	667.8 (1.31)	680.0 (1.65)	666.1 (1.31)
50% censoring	721.2 (3.10)	714.8 (3.04)	722.0 (3.13)
	Expected correct model deviance = 577.2		

To investigate the behavior of the recursive partitioning procedures in a nonproportional hazards setting, H^1 survival times were considered in model D. Table 1 shows that all of the methods do poorly. The FL, MR, and EX methods each selected trees consisting only of the root node in more than 65% of the cases.

In the case of H^1 distributed survival times of model E, the techniques FL and MR still detect structure with high probability; however, the average size of a selected tree is smaller than for the model with proportional hazards, model C. The performance of the EX method is much poorer for uncensored data, with 39% of the trees consisting only of the root node, compared to less than 5% for any of the other methods. The expected H^1 deviance of the EX technique is also substantially larger than for the other methods.

Only model C was considered with a sample size of $n = 500$ observations. Table 2 again shows that the martingale residuals do not perform as well as the other two methods in uncensored data. However, with approximately 50% censoring the MR method has smaller estimated expected deviance than the other two recursive partitioning methods.

In summary, the simulations revealed several differences in performance among the different methods. For uncensored exponential survival times, the martingale residual method did not perform as well as the other methods. Also in the case of exponential survival times, the performance of the one-step full likelihood method was similar to the method based on the correct parametric likelihood, which is expected to perform the best. In addition, the one-step full likelihood method performed substantially better than the parametric method when the data were not exponentially distributed, suggesting that it is an attractive nonparametric alternative.

6. Example

Survival for patients with myeloma, a cancer of the plasma cells of the bone marrow, is strongly associated with age and several laboratory values. In designing clinical trials for myeloma, it is important to protect against possible imbalance of these prognostic factors across treatment groups. Several investigators (Durie and Salmon, 1975; Medical Research Council, 1980) have produced staging or prognostic classification schemes for myeloma that are used for this purpose. However, an unattractive feature of these schemes is that they are based on a variable, tumor mass, that is difficult to determine. A current research goal is to develop a prognostic stratification for myeloma patients based on easily measured variables that could be used as a new staging rule.

This data set is based on 614 patients from a randomized clinical trial for myeloma collected by the Southwest Oncology Group between 1982 and 1987; see Salmon et al. (1990). Five variables previously known to be associated with survival—namely, age, serum calcium, serum albumin, serum creatinine, and serum β_2 microglobulin—were considered for development of a new prognostic stratification or staging system. There are missing values on some covariates for about 70 patients; these observations were included in tree

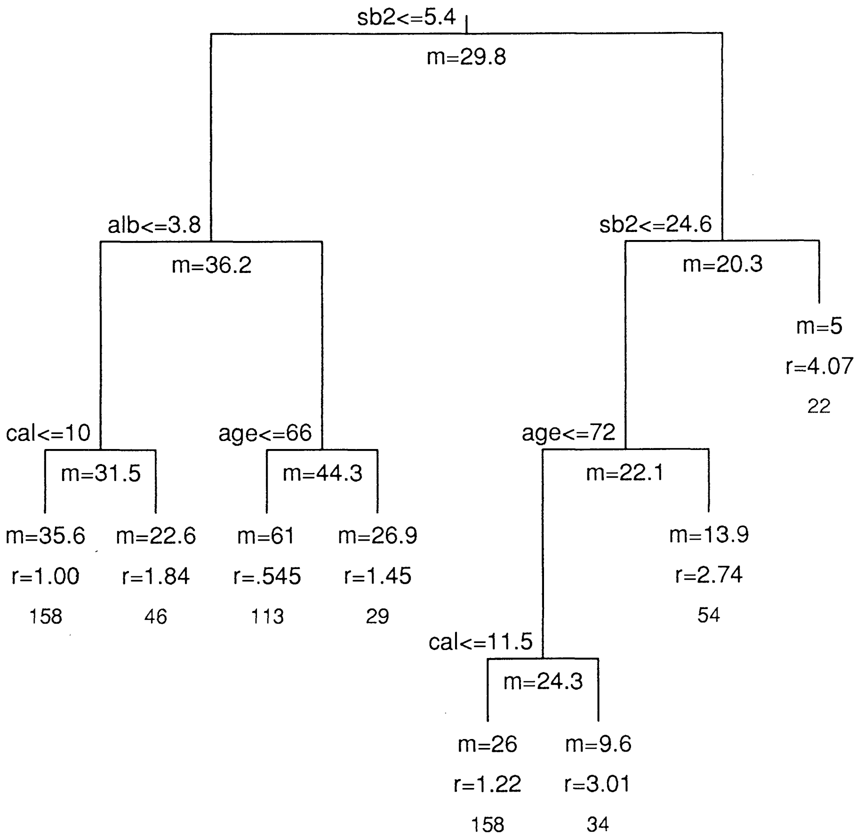

Figure 1. Pruned survival tree for the myeloma data. Median survival in months (m) is given below each node. The relative risk estimate (r), for node h, $\hat{\theta}_h/\hat{\theta}_1$, is given below each terminal node.

analysis by using the surrogate splitting ideas of Breiman et al. (1984). Approximately 34% of the observations are censored.

A tree was grown with a minimum node size of 20 patients. The node size was not chosen to be smaller because extremely small prognostic groups are not of interest for staging and because of the savings in computation time. The one-step full likelihood selected a tree with eight terminal nodes (Figure 1). The covariate and its split value are indicated above each split on the tree. There are splits on four variables; high values of the measurement variables correspond to poorer prognosis except for serum albumin, for which the opposite is true. Kaplan–Meier estimates for each of the terminal nodes are presented in Figure 2.

Recently, there has been discussion in the literature regarding the use of serum β_2 microglobulin for development of staging schemes (Durie et al., 1990; Child et al., 1983; Garwell et al., 1984). This analysis supports the consensus that serum β_2 is an important prognostic factor. The first split in the tree is on serum β_2 microglobulin and divides the data into two groups with median survival of 36 months and 20 months. A small group of patients with high serum β_2 and poor prognosis is also split off.

At this point the investigator may want to reduce the number of groups further. Algorithms for amalgamating the nodes can be constructed and/or the investigators' clinical knowledge may suggest specific groupings. In either case, recursive partitioning seems to be a useful tool for helping the investigator form prognostic groups of patients.

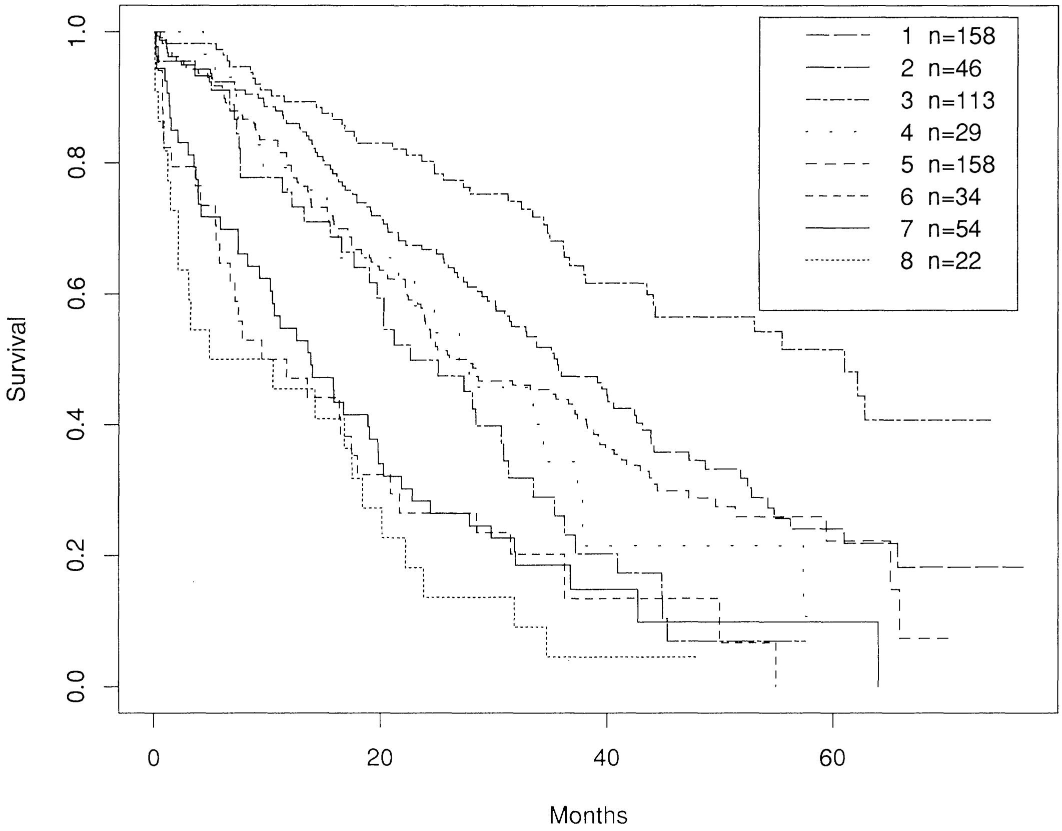

Figure 2. Kaplan–Meier estimates for the terminal nodes given in Figure 1.

7. Discussion

This method of extending the proportional hazards regression to tree-structured relative risk functions performed well on simulated data with proportional hazards survival structure. The method yields interpretable one-step summaries for the terminal nodes of the tree, and full likelihood estimates of the relative risk between observations in nodes can be obtained by an iterative procedure. Weighted least squares scores allow implementation of the algorithm into existing recursive partitioning computer code. The least squares score has not been extensively compared in a simulation study to the one-step full likelihood method; however, in the many examples considered it yielded results similar to those of the one-step full likelihood.

One referee has suggested that the use of the Mantel–Haenszel estimator studied by Crowley, Liu, and Voelkel (1982) may yield an improved tree-based procedure that is still computationally efficient. The use of the Mantel–Haenszel estimator in tree-based methods for survival data should be studied.

We have not described the analytical properties for the pictures and estimates obtained by this procedure. Some consistency results are available for classification and regression trees (Breiman et al., 1984; Gordon and Olshen, 1980, 1984). These results can be extended to show that the estimates of the conditional survival function are also consistent (Butler et al., unpublished technical report, Stanford University, 1989) and that Mantel–Haenszel and Cox-type estimates of the relative risk between a finite number of points in the covariate space converge in probability to the true relative risks if the proportional hazards model is true. We have not yet shown a uniform convergence result for our relative risk estimate. However, there is some question as to the practical relevance of results of this type since they do not apply to the recursive partitioning algorithms as they are almost always implemented. To obtain these consistency results there must also be a sufficient number of observations in the nodes as trees grow. This can be done by considering quantile splits as described by Gordon and Olshen (1984). In addition, the mesh must go to zero; this can be implemented by taking at least some small proportion of the splits to be on each covariate. The latter restriction is an unattractive modification to our highly adaptive regression procedure.

The methods discussed in this paper can be executed fast enough for interactive data analysis since updating algorithms allow splits to be evaluated using $O(N)$ calculations. We believe that interactive tree-structured survival analysis will be a useful adjunct to the tools currently used by those who analyze censored survival data, especially when primary interest is in defining several groups of patients with important differences in survival.

Acknowledgements

The data on patients with multiple myeloma are used with the permission of Dr Sydney Salmon, Chair of the Myeloma Committee of the Southwest Oncology Group. The authors would like to thank Dr Charles A. Coltman, Jr., Chair of the Southwest Oncology Group, for helpful suggestions. This work was supported by National Institutes of Health Grant P01 GM24472 and in part for the first author by a Natural Sciences and Engineering Research Council of Canada Postgraduate Scholarship when he was a graduate student at the University of Washington.

Résumé

Une méthode est développée pour obtenir, à partir de données de survie censurées, des estimateurs du risque relatif déduits d'une structure d'arbre. La première étape d'une procédure d'estimation

reposant sur la vraisemblance complète est utilisée dans un algorithme de partitionnement récursif qui reprend en grande partie l'algorithme, largement utilisé, de Classification et Arbre de Régression (CART) de Breiman et coll. (1984, *Classification and Regression Trees*, Belmont, California: Wadsworth). La performance de la technique est évaluée par simulation, et comparée aux méthodes d'étude de la survie par structure d'arbre proposées par Davis et Anderson (1989, *Statistics in Medicine* **8,** 947–961) et Therneau, Grambsch, et Fleming (1990, *Biometrika* **77,** 147–160).

REFERENCES

Becker, R., Clark, L., and Pregibon, D. (1989). Tree-based models. In *Proceedings of the ASA Statistical Computing Section.* Alexandria, Virginia: American Statistical Association.

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

Breslow, N. (1972). Contribution to the discussion of paper by D. R. Cox. *Journal of the Royal Statistical Society, Series B* **34,** 216–217.

Child, A., Crawford, S., Norfolk, D., O'Quigley, J., Scarfee, J., and Struthers, L. (1983). Evaluation of serum β_2 microglobulin as a prognostic indicator in myelomatosis. *British Journal of Cancer* **47,** 111–114.

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.

Clayton, D. (1983). Fitting a general family of failure-time distributions using GLIM. *Applied Statistics* **32,** 102–109.

Clayton, D. and Cuzick, J. (1985). The EM algorithm for Cox's regression model using GLIM. *Applied Statistics* **32,** 148–156.

Cox, D. R. (1972). Regression models and life tables. *Journal of the Royal Statistical Society, Series B* **34,** 187–200.

Crowley, J., Liu, P. Y., and Voelkel, J. (1982). Estimation of the ratio of hazard functions. *Survival Analysis.* IMS Lecture Notes—Monograph Series #2, 56–73.

Davis, R. and Anderson, J. (1989). Exponential survival trees. *Statistics in Medicine* **8,** 947–961.

Durie, B. and Salmon, S. (1975). A clinical staging rule for myeloma. *Cancer* **36,** 842–854.

Durie, B. G. M., Stock-Novack, D., Salmon, S. E., Finley, P., Beckord, J., Crowley, J., and Coltman, C. A. (1990). Prognostic value of pretreatment serum β_2 microglobulin in myeloma: A Southwest Oncology Group study. *Blood* **75,** 823–830.

Garwell, H., Durie, B., Kyle, R., Findley, D., Bower, B., and Serokman, R. (1984). Serum β_2 microglobulin in the initial staging and subsequent monitoring of monoclonal plasma cell disorders. *Journal of Clinical Oncology* **2,** 51–57.

Gentleman, R. and Crowley, J. (1991). Local full likelihood estimation for the proportional hazards model. *Biometrics* **47,** 1283–1296.

Gordon, L. and Olshen, R. (1980). Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis* **10,** 611–627.

Gordon, L. and Olshen, R. (1984). Almost surely consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis* **15,** 147–163.

Gordon, L. and Olshen, R. (1985). Tree-structured survival analysis. *Cancer Treatment Reports* **69,** 1065–1069.

Harrington, D. and Fleming, T. (1982). A class of rank test procedures for censored survival data. *Biometrika* **69**, 553–566.

LeBlanc, M. (1990). Tree-based tools for survival data. In *Proceedings of the XV International Biometrics Conference*, 123–138. Alexandria, Virginia: The Biometric Society.

McCullagh, P. and Nelder, J. (1983). *Generalized Linear Models.* London: Chapman and Hall.

Medical Research Council's Working Party on Leukemia in Adults. (1980). Prognostic features in the third MRC myelomatosis trial. *British Journal of Cancer* **42,** 831–840.

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 vector when only the coordinate is observable. *Technometrics* **12,** 923–924.

O'Sullivan, F. (1988). Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal of Scientific and Statistical Computing* **9,** 531–542.

Salmon, S. E., Tesh, D., Crowley, J., Saeed, S., Finley, P., Milder, M. S., Hutchins, L. F., Coltman,

C. A. Jr., Bonnet, J. D., Cheson, B., Knost, J. A., Samhouri, A., Beckord, J., and Stock-Novack, D. (1990). Chemotherapy is superior to sequential hemibody irradiation for remission consolidation in multiple myeloma: A Southwest Oncology Group study. *Journal of Clinical Oncology* **8,** 1575–1584.

Segal, M. (1988). Regression trees for censored data. *Biometrics* **44,** 35–48.

Therneau, T., Grambsch, P., and Fleming, T. (1990). Martingale based residuals for survival models. *Biometrika* **77,** 147–160.

Tibshirani, R. and Hastie, T. (1987). Local likelihood estimation. *Journal of American Statistical Association* **82,** 559–567.

Received March 1990; *revised November* 1990; *accepted January* 1991.