*Statistical Methods in Medical Research* 1995; **4**: 237–261

# Trees and splines in survival analysis

**Orna Intrator** Department of Statistics, Hebrew University, Jerusalem, Israel and
**Charles Kooperberg** Department of Statistics, University of Washington, Seattle, USA

During the past few years several nonparametric alternatives to the Cox proportional hazards model have appeared in the literature. These methods extend techniques that are well known from regression analysis to the analysis of censored survival data. In this paper we discuss methods based on (partition) trees and (polynomial) splines, analyse two datasets using both Survival Trees and HARE, and compare the strengths and weaknesses of the two methods. One of the strengths of HARE is that its model fitting procedure has an implicit check for proportionality of the underlying hazards model. It also provides an explicit model for the conditional hazards function, which makes it very convenient to obtain graphical summaries. On the other hand, the tree-based methods automatically partition a dataset into groups of cases that are similar in survival history. Results obtained by survival trees and HARE are often complementary. Trees and splines in survival analysis should provide the data analyst with two useful tools when analysing survival data.

## 1 Introduction

In this paper we discuss and compare two groups of nonparametric methodologies for the analysis of censored survival data, methods based on recursive partitioning and those based on polynomial splines. Traditional methods for analysing survival data include exploratory methods such as the Kaplan–Meier estimate, the Nelson–Aalen estimate, and various types of tests that summarize differences between two or more survival distributions, and modelling methods such as the Cox proportional hazards model and the accelerated lifetime model. The nonparametric methods discussed in this paper can give insight into data that the traditional methods fail to provide.

The use of *Classification and regression trees*[1] (CART) and other recursive partitioning methods have allowed a more thorough examination of effects of variables on the survival distribution. When modelling survival data, it is frequently of interest to determine which variables affect the survival distribution and whether the effect is valid across all individuals or within subsets. Statistically, these questions are often posed as variable selection and detection of interactions. CART is used as a tool for revealing structure in the data. It is the ease of interpretation of the results and the ability to analyse complex nonlinear datasets with many variables by effectively reducing the dimensionality of the data that are CART's main advantages over other methods. All these features are highly desirable for exploratory analyses.

Polynomial splines form a versatile tool for function estimation. They have been used in many situations such as multiple regression,[2] density estimation,[3] estimation of the spectral distribution,[4] and polychotomous regression and classification.[5] In the polynomial spline approach, an unknown function is modelled in a linear space. Stepwise algorithms make it possible to determine this space adaptively. In the proportional hazards model[6] the conditional log-hazard function is an additive function of time and the vector of covariates. Traditionally, in this model the dependence of the survival time on the covariates is modelled fully parametrically, so

Address for correspondence: C Kooperberg, Department of Statistics, Box 354322, University of Washington, Seattle, WA 98195-4322, USA.

that the regression function can be estimated independently of the baseline hazard function. When using polynomial splines, such parametric assumptions are not needed. It also becomes convenient to fit and compare linear proportional hazards models, additive proportional hazards models, proportional hazards models with time-varying coefficients, and nonparametric proportional hazards models.

This paper is organized as follows: in Sections 2 and 3 we give overviews of the use of trees and splines in survival analysis. In particular, we discuss survival trees (in Sections 2.3 and 2.4) and HARE (in Section 3.2) in some detail. In Section 4 we apply both methods to two examples. We end the paper by discussing the strengths and weaknesses of the two methods.

## 2 Trees in survival analysis

### 2.1 Overview

Several extensions of CART to censored survival data, sometimes termed survival trees, have been proposed in the literature. (See Section 2.2 below for a brief review of CART.) These extensions typically require modification of the four basic building blocks of CART: the prediction rule, the splitting rule, the pruning algorithm and the tree selection. The prediction rule for survival analysis is typically based on the estimate of the distribution function, which implies that the three other ingredients all work nonparametrically on the space of distributions. Applications of survival trees may be found in several articles.[7–12]

The extensions of CART to survival data fall into two groups. One approach uses a statistic that determines within-node homogeneity: how similar are the survival experiences of observations in a node.[13–17] The alternative approach is based on separation measures. The main ingredient is now a (test) statistic that distinguishes between survival experiences.[18–23]

Gordon and Olshen[13] presented the first extension of CART to censored survival data, which involved a distance measure (the Wasserstein metric) between Kaplan–Meier curves and certain point masses. Their approach amounts to assuming a piecewise exponential model with one data-determined knot. Ciampi *et al.*'s method[14] is based on a parametric model and likelihood ratio statistics. David and Anderson[15] suggest a method based on the observed likelihood at a node, while assuming an exponential model for the baseline hazard function. LeBlanc and Crowley[16] use deviance residuals based on the Cox proportional hazards model for the splitting rule. These extensions of CART are based on a definition of a within-node homogeneity measure. The use of within-node homogeneity based on likelihood statistics allows the inheritance of all subsequent CART methodology, since the measures defined are all subadditive, allowing comparison between subtrees. However, except for Gordon and Olshen's work, these methods are not devoid of parametric assumptions about the underlying hazards model. Zhang[17] uses a within-node homogeneity measure that is based on the idea that a homogeneous node should consist of subjects whose observed failure times are close and who are mostly censored or mostly uncensored.

Segal[19] argued that tests for between-node separation can tell more about the important prognostic factors associated with the survival phenomenon under study than within-node homogeneity. He introduced a totally nonparametric tree algorithm, basing the partitioning on between-node separation using the Harrington–Fleming[24]

class of two sample rank statistics. Pruning based on between-node separation, as discussed in,[18,19,22] is conceptually harder.

## 2.2 Review of CART

The Classification and Regression Trees (CART) method of Breiman *et al.*[1] addresses the classification and regression problem by building a binary decision tree according to some splitting rule based on the covariates. In this way, the space of explanatory variables is partitioned recursively in a binary fashion. The partitioning is intended to increase within-node homogeneity, where homogeneity is determined by the dependent variable in the problem. The partitioning is repeated until a node is reached for which no split improves the homogeneity, whereupon the splitting is stopped and this node becomes a terminal node. Prediction is determined by terminal nodes, and is either a class label in classification problems, or the average of the dependent variable in least-squares regression problems.

A tree $T$ has a root node whose subnodes (also called daughters) can be divided into terminal nodes, collectively denoted by $\tilde{T}$, and decision nodes. The number of terminal nodes is denoted by $|\tilde{T}|$. The branch $T_t$ that stems from node $t$ includes $t$ itself and all its subnodes. This branch has terminal nodes collectively denoted $\tilde{T}_t$.

In the regression context let $Y_i$ be the dependent variable and $\mathbf{x}_i$ the vector of covariates for the $i$th observation, $1 \leq i \leq N$. Least squares regression trees look for a predictor with constant value at each node $t$ for which the splitting rule $s_t^*$ maximizes the difference of the weighted squared error between $t$ and its daughters $t_L$ and $t_R$ over the set $S$ of all possible splits; that is.

$$\Delta R(t,s_t^*) = \max_{s \in S}[R(t) - R(t_L) - R(t_R)],$$

where the risk at the node $t$ is given by $R(t) = p(t)s^2(t)$ with $p(t) = N(t)/N$ being the proportion of observations that fall in $t$ and $s^2(t)$ being the variance of the corresponding values of the dependent variable.

A useful feature of CART is that of growing a large tree and then pruning it to get a sequence of nested pruned subtrees. This is done using a penalty measure based on the complexity of the tree and called the cost complexity. Thus

$$R_a(T) = R(T) + a|\tilde{T}|, \tag{2.1}$$

where $R(T) = \sum_{t \in \tilde{T}} R(t)$ and $a$ is a penalty parameter for the complexity of the tree.

A sequence of nested trees is constructed in a bottom-up approach, starting with the fully grown tree, which corresponds to the smallest value of $a$ and with smaller nested subtrees corresponding to increasing values of $a$. This is done by recursively pruning the branch(es) with the weakest link; that is, the node $t$ with the smallest value of $a$ such that $R_a(t) \leq R_a(T_t)$.

In least squares regression trees, $R(t) - R(T_t) \geq 0$. This subadditivity of the risk $R(t)$ is of utmost importance in pruning in that it allows the comparison of different branches of the tree as well as the definition of nested subtrees.

To select the best pruned subtree, CART suggests finding honest estimates of the error $R(T)$ using a test sample or crossvalidation. In regression the error denoted by $R(T)$ is the average squared error in the terminal nodes. Tree selection can also be done in an exploratory manner by examining trees in sequence.

### 2.3 Survival trees based on between-node separation

Segal[19], Intrator[18] and LeBlanc and Crowley[22] use as the prediction rule the Kaplan–Meier estimate of the survival distribution, and as the splitting rule a test for measuring differences between distributions adapted to censored data such as the logrank test or Wilcoxon test, and more generally the $G^{\rho,\gamma}$ class of rank statistics.[25,26] These statistics are weighted versions of the logrank statistic, where the weights allow flexibility in emphasizing differences between two survival curves for early times (the left tail of the distribution), middle times or late times (the right tail of the distribution). In particular, an observation at time $t$ is weighted by $Q(t) = \hat{S}(t)^{\rho}(1 - \hat{S}(t))^{\gamma}$, where $\hat{S}$ is the Kaplan–Meier estimate of the survival curve for both samples combined. Thus we can obtain sensitivity to early occurring differences by taking $\rho > 0$ and $\gamma \approx 0$, while we emphasize differences in the middle by taking $\rho \approx 1$ and $\gamma \approx 1$ and we emphasize late differences if $\rho \approx 0$ and $\gamma > 0$. The traditional logrank statistic is obtained when $\rho = \gamma = 0$.

As mentioned in the previous sections, it is important that a risk measure, such as $R(t)$, is monotonic. In particular, for cost complexity pruning, the improvement in node homogeneity or a measure of the quality of a branch relative to its root needs to be measured. Once a measure $m(t,T_t)$ of the quality of a branch is defined, it can be used in cost complexity pruning by comparing it with the number of terminal nodes in the branch $|\tilde{T}t|$. The quality should be monotonically decreasing down the decision nodes of the tree. In Intrator's version of survival trees, the definition of $m(t,T_t)$ is based on the significance level. Let $v(t)$ be one minus the $p$ value of a test statistic for difference between survival distributions for $t_L$ and $t_R$. Define the *quality* of a decision node $t$ by

$$m(t) = p(t) \max_{\tau \in T_t - \tilde{T}_t} v(\tau),$$

where $p(t)$ is the probability of being in node $t$. The quality $m(t)$ measures how much the risk decreases if we use the best possible split on node $t$, relative to not splitting node $t$ at all. Segal[19] used a similar measure of quality $m(t)$, the maximal chi-squared statistic of the branch $T_t$, but he did so without weighting by the probability of the node.

In all of these measures, the maximization is done over all the decision nodes of the branch $T_t$. This results in monotonicity of the quality down the tree, which is necessary for defining effective nested pruning.

Cost-complexity pruning is possible by comparing $m(t)$ with the number of terminal nodes: $m(t) = ag(|\tilde{T}_t|)$ where $g(x)$ is some monotonically nondecreasing function. For example, Intrator[18] chooses $g(x) = x - 1$. We get an increasing sequence of the critical pruning values: for tree $T_{j-1}$ in the sequence we define $a_j$ as $a_j = \min_{t \in Tj} m(t)/g(|\tilde{T}_t|)$. Tree $T_j$ is then tree $T_{j-1}$ pruned at level $a_j$. The sequence of $a$s defines a nested sequence of pruned subtrees.

As in CART, exploratory tree selection can be done by examining the plot of values of the penalty parameter $a$ versus tree size.

Tree selection may be based on goodness-of-prediction measures. Intrator[18] explores the measure given by

$$\mathrm{PE}(T_a) = \sum_{t \in \tilde{T}_a} p(t) v(S^{ts}(t), S^{ls}(t)),$$

where $S^{ts}(t)$ is the estimated survival curve at node $t$ based on a test sample, $S^{ls}(t)$ is

that based on the learning sample, and $1 - v(\cdot,\cdot)$ is the $p$ value of a test statistic for the difference between survival distributions. Estimation of goodness-of-prediction may also be done using a crossvalidation scheme. Another approach is to validate the complexity parameters. The idea here is to test the credibility of the pruning and not the goodness-of-prediction directly. This approach is currently being investigated by Intrator.

Alternatively, LeBlanc and Crowley[22] measure the quality of a branch $T_t$ as the sum of the logrank test statistic values (chi-squared values) of the decision nodes of the branch, denoted by $G(T_t)$. They define *split complexity pruning* as $G_a(T_t) = G(T_t) - a|T_t - \tilde{T}_t|$, i.e. their complexity is based on the number of decision nodes in the branch. LeBlanc and Crowley compute their split complexity measure (based on their quality measure) on test data, using a fixed penalty term $a$ which is usually selected to be in the range of two to four or they employ resampling techniques.

Many other useful features of CART can be incorporated in survival trees. For example, Intrator[18] explores tree robustness using a crossvalidation approach while growing the tree, testing all possible splits on several samples and selecting the split that was the best on most samples. The uses of surrogate splits for handling missing data and variable importance, ideas initially presented in CART,[1] are also extended to the survival trees setting.

### 2.4 Survival trees based on within-node homogeneity

Tree building and pruning based on within-node homogeneity allows for trivial inheritance of the CART algorithm. Often, tree growing is based on a between-node separation measure, while tree pruning, and selection are based on a within-node homogeneity measure. Once a split has been determined based on a between-node separation measure, the statistic measuring within-node homogeneity can be evaluated at the daughter nodes for further use in pruning, as done in CART for classification.

Gordon and Olshen's within-node homogeneity measure is the $L_p$ Wassertstein distance between the Kaplan–Meier estimate of survival in a node and a survival curve defined by a piecewise constant hazard function with a data-determined single point of discontinuity. When $L_2$ Wasserstein distances are used, the homogeneity corresponds to the variance of the Kaplan–Meier estimate.

Davis and Anderson[15] define within-node homogeneity based on the negative log-likelihood of an exponential model at the node. Their measure for homogeneity of node $t$ is $-\hat{\ell}(t) = D_t - D_t\log(D_t/Y_t)$ where $D_t$ is the number of failure events in the node and $Y_t$ is the total observation time (time on test) in the node.

LeBlanc and Crowley[16] assume a semiparametric proportional hazard model, where the hazard $\lambda(t|z_i)$ at time $t$ for individual $i$ with covariates $z_i$ is the product of a baseline hazard that depends only on time and a structural component that depends on the individual through its covariates $\lambda_0(t)\theta(z_i)$. Their within-node homogeneity measure is based on a single step deviance residual. The deviance residual for individual $i$ is defined as $2[\ell_i(\text{saturated}) - \ell_i(\hat{\theta}_{MLE})]$ where the log-likelihood $\ell_i(\text{saturated})$ corresponds to a saturated model, which allows a parameter for each individual and $\ell_i(\hat{\theta}_{MLE})$ corresponds to the maximum likelihood estimate for the present tree based on the proportional hazards model. LeBlanc and Crowley use a single step estimate of the deviance, which is based on the Breslow[27] estimate of the baseline hazard using the Nelson[28] estimate of the structural component (which is one for all individuals). The resulting deviance residual for individual $i$ is given by

$$d_i = 2[\delta_i \log(\delta_i / \Lambda_0(t_i)) - (\delta_i - \Lambda_0(t_i))].$$

The baseline cumulative hazard at node $k$ is given by the Breslow estimate

$$\Lambda_0(t) = \sum_{i:t_i \leq t} \delta_i \Big/ \left( \sum_{i \in k} \sum_{i:t_i > t} 1 \right).$$

LeBlanc and Crowley show that this impurity measure 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.

Lastly, Zhang[17] introduces a totally different concept of splitting. He argues that a homogeneous node should consist of subjects whose observed failure times are close and who are mostly censored or mostly uncensored. Thus he suggests a splitting criterion based on a definition of node impurity which is a weighted combination of impurity of the censoring and of the times. The impurity is thus a combination of CART's original impurity measure for classification (of the censoring indicators) and of regression (MSE). In his paper he discusses the merits of this splitting method, and compares it with Segal's method, Davis and Anderson's method and Gordon and Olshen's method on simulated data. Zhang does not consider pruning or tree selection: he believes that the pruning should be done manually from the full tree by a practitioner in the field of application with the guidance of the computer output (personal communication).

## 3 Splines in survival analysis

### 3.1 Overview

The use of splines has led to a number of new methodologies for survival analysis. These methods roughly divide into two groups: those that make use of penalized likelihood estimation, which we refer to as smoothing splines methods, and those that use polynomial splines often in conjunction with adaptive model selection.

The smoothing spline solution to a function estimation problem is typically the maximizer of a penalized likelihood function.[29–31] In survival analysis, smoothing splines have been used by Anderson and Senthilselvan,[32] Whittemore and Keller,[33] Senthilselvan,[34] O'Sullivan,[35,36] Gray,[37] Hastie and Tibshirani[30,38] and Gu.[39,40] Most of these papers use splines within the framework of the proportional hazards model. Gray[37] and Hastie and Tibshirani[38] use time-varying coefficients. Gu[40] models the complete log-hazard function, though the computational demands seem too formidable to be applicable in situations with many covariates such as our examples in Section 4. To make the computations more feasible, Gray[37] uses a B-spline approximation to the smoothing spline problem.

In the polynomial spline approach, an unknown function is modelled in a linear space. Stepwise algorithms make it possible to determine this space more adaptively than in the smoothing spline approach, which involves only a few smoothing parameters. Adaptive algorithms for polynomial splines were first introduced in a regression context.[41] Other applications include multiple regression (MARS),[2] density estimation (LOGSPLINE),[3] and spectral distribution estimation (LSPEC).[4]

In the context of survival analysis, Etezadi-Amoli and Ciampi,[42] Efron[43] and Abrahamowicz, Ciampi and Ramsay[44] use polynomial splines to model either the unconditional distribution of the survival times or the baseline hazard function within

a proportional hazards model. Kooperberg, Stone and Truong[45] develop hazard regression (HARE), in which the conditional log-hazard function is modelled using polynomial splines. Kooperberg[46] extends the HARE methodology to handle interval-censored data. Under suitable conditions Kooperberg, Stone and Truong[47] obtain the $L_2$ rate of convergence for a nonadaptive version of HARE.

### 3.2 HARE

In hazard regression (HARE, Kooperberg, Stone and Truong,[45] hereafter referred to as KST), polynomial splines are used to estimate the conditional log-hazard function based on possibly censored survival data and one or more covariates. An automatic procedure involving maximum likelihood, stepwise addition, stepwise deletion and BIC is used to select the final model. The possible models contain proportional hazards models as a subclass, which makes it possible to diagnose departures from proportionality. We now summarize the HARE algorithm.

Let $T$ be a (nonnegative) survival time whose distribution may depend on a vector of $M$ covariates $\mathbf{x} = (x_1,\ldots,x_M)$ ranging over a subset $X = X_1 \times \ldots \times X_M$ of $\mathrm{R}^M$. Let $f(t|\mathbf{x})$, $F(t|\mathbf{x}) = \int_0^t f(u|\mathbf{x})du$, $\lambda(t|\mathbf{x}) = \mathrm{f}(t|\mathbf{x})/[1 - \mathrm{F}(t|\mathbf{x})]$ and $\alpha(t|\mathbf{x}) = \log\lambda(t|\mathbf{x})$ denote the corresponding conditional density, distribution, hazard and log-hazard functions, respectively.

Let $G$ be a $p$-dimensional linear space of functions on $[0,\infty) \times X$, and let $B_1,\ldots,B_p$ be a basis of $G$. The HARE model for $\alpha(t|\mathbf{x})$ is given by

$$\alpha(t|\mathbf{x};\beta) = \sum_{j=1}^{p} \beta_j B_j(t|\mathbf{x}). \qquad t \geq 0. \tag{3.1}$$

The basis functions of $G$ that HARE allows are piecewise linear functions (splines) in the covariates, piecewise linear functions in $t$, and tensor products of two such piecewise linear functions. (The tensor product of the functions $g_1(x_1)$ and $g_2(x_2)$ is the function $g_1(x_1)g_2(x_2)$.) Both the space $G$ and its dimension $p$ are determined adaptively. To use HARE models we have to resolve two issues: (1) how to select $p$ and $G$; and (2) how to estimate $\beta_1,\beta_2,\ldots,\beta_p$ given $G$.

Before pursuing these issues further, we should point out that if none of the basis functions of $G$ depend on both $t$ and $\mathbf{x}$, then (3.1) is a proportional hazards model.[6] It is a particular interesting feature of HARE that the model selection procedure may or may not result in such a model. If any of the basis functions in the selected model is a tensor product of a piecewise linear function in $t$ and a piecewise linear function in one of the covariates, then a proportional hazards model might not be appropriate.

Given a $p$-dimensional linear space $G$ with basis $B_1,\ldots,B_p$, the coefficients $\beta_1,\beta_2,\ldots,\beta_p$ can be estimated by maximum likelihood. In particular, consider $n$ randomly selected individuals. For $1 \leq i \leq n$, let $T_i$ be the survival time, $C_i$ the censoring time, and $\mathbf{x}_i$ the vector of covariates for the $i$th such individual, and set $Y_i = \min(T_i,C_i)$ and $\delta_i = \mathrm{ind}(T_i \leq C_i)$. The random variable $Y_i$ is said to be uncensored or censored according as $\delta_i = 1$ or $\delta_i = 0$. Note that the partial likelihood corresponding to $Y_i = y_i$, $\delta_i$, $\mathbf{x}_i$ and $\beta$ equals $[f(y_i|\mathbf{x}_i;\beta)]^{\delta_i}[1 - F(y_i|\mathbf{x}_i;\beta)]^{1-\delta_i}$ (Miller, p 16),[48] so the log-likelihood is given by

$$\phi(y_i,\delta_i|\mathbf{x}_i;\beta) = \delta_i\alpha(y_i|\mathbf{x}_i;\beta) - \int_0^{y_i} \exp(\alpha(u|\mathbf{x}_i;\beta))du, \qquad y_i \geq 0 \text{ and } \delta_i \in \{0,1\}.$$

The log-likelihood function corresponding to the observed data is given by

$$l(\beta) = \sum_i \phi(Y_i,\delta_i|\mathbf{x}_i;\beta), \qquad \beta \in \mathbf{R}^p.$$

It is straightforward to compute the corresponding score $\mathbf{S}(\beta)$ and Hessian $\mathbf{H}(\beta)$. In particular, it is easily established that the log-likelihood function is concave. The maximum likelihood estimate $\hat{\beta} = \arg\max_{\beta} l(\beta)$ can thus be found using the Newton-Raphson algorithm, and the log-likelihood of the fitted model is given by $\hat{l} = l(\hat{\beta})$.

The selection of $p$ and $G$ is carried out using an algorithm that employs stepwise addition and stepwise deletion of basis functions. Initially we fit the one-dimensional model $\alpha(t|\mathbf{x};\beta) = \beta_1$. Then we proceed with stepwise addition. Here we successively replace the $(p-1)$-dimensional space $G_0$ by a $p$-dimensional space $G$ containing $G_0$ as a subspace. Since it is computationally too time consuming to evaluate each candidate for a new basis function by recomputing the maximum likelihood fit, we choose among the various candidates by a heuristic search that is designed approximately to maximize the absolute value of the corresponding Rao statistic.[49] This is similar to what is sometimes done for generalized linear models; see, for example, the function `step.glm` in S and the discussion in Chambers and Hastie (p 235).[50] (After the new basis function has been added to the model, it is refitted using maximum likelihood and the Newton-Raphson algorithm.) As mentioned earlier, the candidate basis functions of $G$ are piecewise linear functions (splines) in the covariates, piecewise linear functions in $t$, and tensor products of two of such piecewise linear functions. However, because of regularity conditions on $G$, not all potential basis functions can be added at any time; for example, tensor products involving a basis function can only be added to the model if the basis function itself has already been added. See KST for more details.

Upon stopping the stepwise addition stage, we proceed to stepwise deletion. Here we successively replace the $p$-dimensional space $G$ by a $(p-1)$-dimensional subspace $G_0$ until we arrive at a one-dimensional space, at each step choosing the candidate space $G_0$ so that the Wald statistic for a basis function that is in $G$ but not in $G_0$ is smallest in magnitude. As in stepwise addition, we do not refit the model for each basis function that is a candidate to be dropped.

During the combination of stepwise addition and stepwise deletion, we get a sequence of models indexed by $\nu$, with the $\nu$th model having $p_\nu$ parameters. Let $\hat{l}_\nu$ denote the log-likelihood of the $\nu$th model, and let

$$\mathrm{AIC}_{a,\nu} = -2\hat{l}_\nu + ap_\nu \tag{3.2}$$

be the Akaike information criterion with penalty parameter $a$ for this model. In this paper we will use $a = \log n$ as in the Bayesian information criterion (BIC). We select the model corresponding to the value $\hat{\nu}$ of $\nu$ that minimizes $\mathrm{BIC}_\nu = \mathrm{AIC}_{\log n,\nu}$.

A program for implementing HARE has been written in C, and an interface based on the statistical package S[50,51] has also been developed. For a more detailed discussion of the HARE procedure and its interface, see KST.

There are a number of similarities between the survival tree algorithm and the HARE algorithm:

1) The cost complexity in survival trees (equation 2.1) corresponds to AIC in HARE (equation 3.2).
2) The splitting algorithm in survival trees is similar to the stepwise addition in HARE. (The heuristic search to maximize the Rao statistic in HARE addresses a problem that is similar to one in survival trees: where to locate the next cutpoint.)

3) The pruning algorithm in survival trees can be compared to the stepwise deletion algorithm in HARE. (Note, however, that in survival trees several nodes may be pruned at the same time, while in HARE basis functions are removed one at a time.)

## 4 Examples

### 4.1 Breast cancer data

The data for our first example come from six breast cancer studies conducted by the Eastern Cooperative Oncology Group. It has been analysed in Gray[37] using a hybrid of penalized likelihood and polynomial splines (see Section 3) and in KST using HARE (see Section 3.1). In this subsection we present a survival tree analysis, followed by a summary and extension of the HARE analysis. We end the subsection with a proportional hazards analysis and a comparison of the three methods.

There were 2404 breast cancer patients in the six studies. All patients had disease involvement in their axillary lymph nodes at diagnosis indicating some likelihood that the cancer had spread through the lymphatic system to other parts of the body; however, none of the patients had evidence of disease at the time of entry into the study, which was following surgical removal of the primary tumour and axillary metastases. The response is survival time (years) from entry into the study. There are six covariates, oestrogen receptor status (ER: 0 is 'negative', 1 is 'positive'), the number of positive axillary lymph nodes at diagnosis, size of the primary tumour (in mm), age at entry, menopause (0 is premenopause, 1 is postmenopause), and body mass index (BMI: defined as weight/height$^2$ in kg/m$^2$). Since the empirical distribution of the number of nodes is highly skewed to the right, we used log(number of nodes) instead of the number itself in the HARE analysis. Of the 2404 cases, 1116 were uncensored and 1288 were censored. There were no missing values for any of the covariates.

*Survival trees analysis*

In Sections 2.3 and 2.4 we briefly described a number of survival tree methods.[13,15,16,18,19,22] In the examples section we mainly present the results from Intrator's method,[18] and briefly describe results from Davis and Anderson's method[15] and LeBlanc and Crowley's method.[22]

Variables analysed in survival trees are always reduced to dichotomies. The present version of the program developed by Intrator does not automatically test all possible break points for continuous variables, and a set of binary variables representing ranks must be provided to it.

In the breast cancer data there are two dichotomous variables: ER and menopause, and four continuous variables: number of nodes, size of the primary tumour, BMI and age. Since the tree looks for binary splits, monotone transformation of the variables (such as taking the log of the number of nodes as in the HARE analysis) is irrelevant. The splits examined for number of nodes were less than or equal to $k$ versus greater than $k$, for $k = 1,2,\ldots,7$. The possible split points for size were 10 mm, 15 mm,. . .,65 mm. The possible BMI split points were 20, 22.5, 25, 29 and 33 kg/m$^2$. The age split points examined were 30, 40, 50, 60 and 70 years.

The survival trees program of Intrator[18] was run using all splits, a partial set of splits, and on different randomly selected subsamples to test for tree robustness and

competing structures. A variety of pairs of values for $\rho$ and $\gamma$ were used for emphasizing early, middle and late differences between survival experiences. The tree selection was carried out in an exploratory manner after examining the prediction error. The trees appeared sensitive to changes in the splitting rule, suggesting a complex structure and perhaps a more complicated nonlinear model.

The first split was always on the number of nodes, with split points usually at four to six nodes. The branch with the smaller number of nodes was usually split on size, at cutpoints between 35 and 50 mm, while the branch with the larger number of nodes was typically split on ER. Age, BMI and menopausal state all occurred further down in some trees, but there was no clear picture. This suggests that there may be no further well defined strata, but that these variables are useful as predictors in a model that is not easily described by a partition tree.

In Figure 1 we display a tree with nine terminal nodes obtained using the splitting rule with $\rho = 1$ and $\gamma = 0.1$, which emphasizes early differences. In Table 1 we show the relation between the cost-complexity parameter $a$, the goodness-of-prediction statistic $PE(T_a)$ and the number of terminal nodes $|\tilde{T}_t|$. Note that we get a tree with nine terminal nodes with any choice of $a$ between 0.162 and 0.277, while the range of values for which we get a tree with eight terminal nodes or ten terminal nodes is much smaller. We also note from this table that $PE(T_a)$ starts to level off at about nine terminal nodes. Considering cost complexity and prediction error together we decided on a tree with nine nodes. Typically we expect a prediction error based on crossvalidation to reach a minimum value, which has not yet happened in Table 1. One reason why the prediction error might not have reached its minimum value is that an accurate description of the data may require a larger but less interpretable tree. Another possible reason is that PE is artificially reduced by the small numbers of cases in the terminal nodes of the larger trees.

The next (tenth) split was a split of terminal node IV on ER and the eleventh split was a split of terminal node VIII on BMI. There are two consecutive splits in the tree on nodes, which effectively form a three-way split. Similarly there are two consecutive splits on size of the group with four or less nodes. Note also that the tree has a four-way interaction involving nodes, size, BMI and menopausal status and a three-way interaction involving nodes, ER and menopausal status.

In Figure 2 we show estimates for the survival functions for the nine terminal nodes. We show both the usual Kaplan–Meier curves (left side) as well as estimates using Hazard Estimation with Flexible Tails[45] (HEFT). HEFT employs cubic splines and has some additional log terms that make it possible to estimate tails more flexibly; otherwise HEFT is very similar to HARE. See KST for more details.

Note that many of the survival functions cross each other. This suggests that a proportional hazards model may not be appropriate for the data, which we will also conclude in the HARE analysis below. We also notice that some curves, like those

**Table 1** Cost complexity and tree size for the breast cancer data

| Number of terminal nodes $[\tilde{T}]$ | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| Cost complexity $a$ | 1.000 | 1.000 | 0.723 | 0.592 | 0.553 | 0.445 | 0.307 |
| Goodness-of-prediction $PE(T_a)$ | 0.558 | 0.427 | 0.329 | 0.316 | 0.244 | 0.231 | 0.215 |
| Number of terminal nodes $[\tilde{T}]$ | 8 | 9 | 10 | 11 | 12 | 13 | 14 |
| Cost complexity $a$ | 0.277 | 0.162 | 0.131 | 0.123 | 0.113 | 0.097 | 0.090 |
| Goodness-of-prediction $PE(T_a)$ | 0.207 | 0.201 | 0.187 | 0.174 | 0.173 | 0.157 | 0.156 |

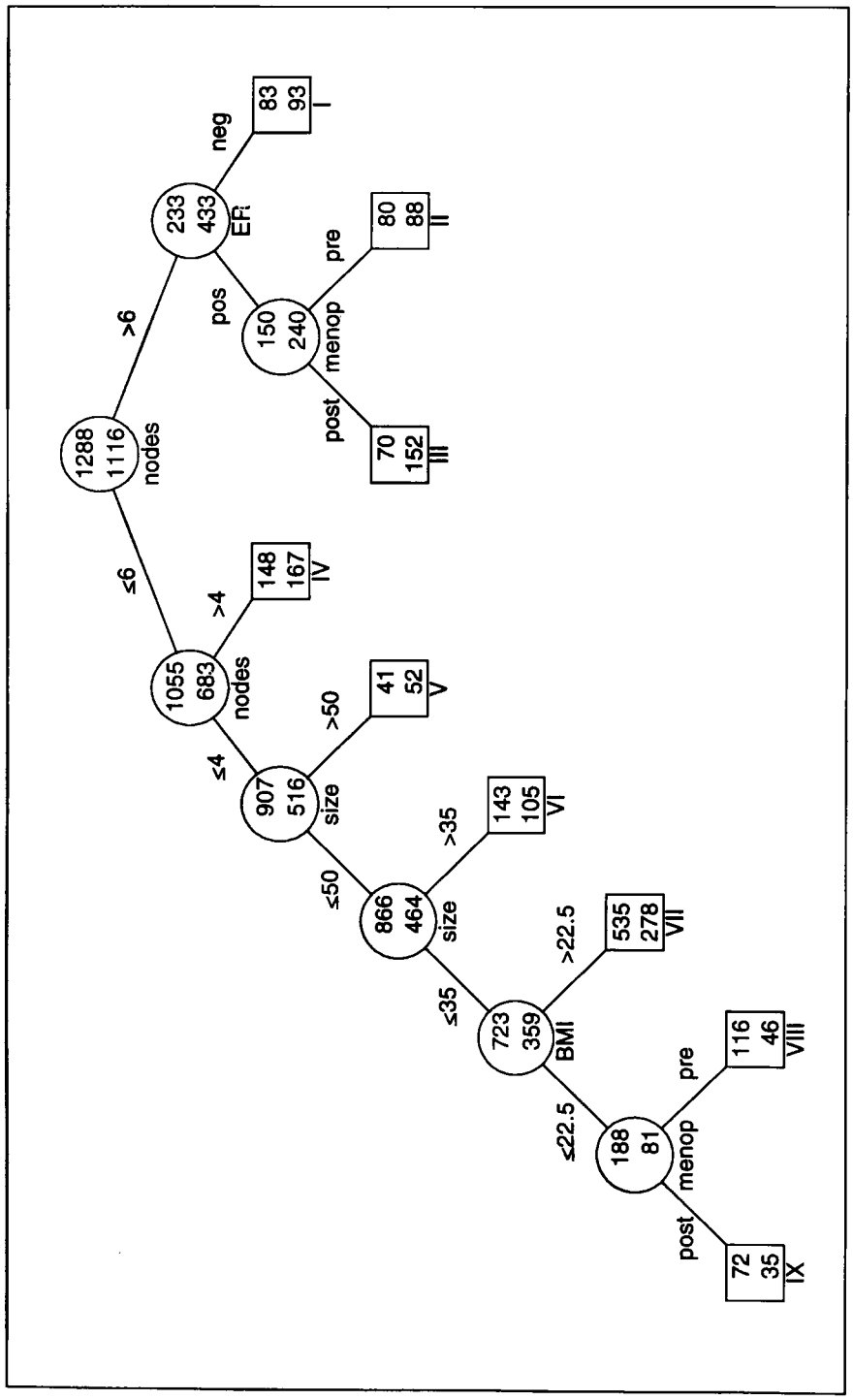


**Figure 1** Survival tree for the breast cancer data. Terminal nodes are indicated by squares and nonterminal nodes by circles. Numbers in circles and squares indicate the number of alive (top) and deceased (bottom) cases in that node.

corresponding to terminal nodes IV and VII, cluster together, although their nodes were in different parts of the tree. Thus our analysis, combining survival trees and HEFT, enable us to identify different groups of people with similar survival experiences.

The tree obtained using the method of LeBlanc and Crowley[22] was very similar to the tree we presented in Figure 1. It also first split on nodes at (nodes $\leq 6$). After which the left side was split at (nodes $\leq 3$) and the right side was split on ER. The two next splits were also very similar: the node with (nodes $\leq 3$) was split at (size $\leq 52$) and the node with ER positive was split on menopausal state. Splits further down the tree were more different.

The tree obtained using the method of Davis and Anderson[15] split at the root node to the left on (nodes $\geq 4$), after which the left node split on (nodes $\leq 6$) and the right node split on (size $\leq 25$). This presents a model similar to that presented in Figure 1, where the smaller number of nodes is split on size. However, the structure is much smaller.

*HARE analysis*

Before applying HARE it is often advantageous to use (unconditional) hazard estimation to transform time so that the transformed unconditional hazard function will be approximately equal to one.[52] The main advantage of such a transformation is that because of the piecewise linear nature of HARE, the (baseline) hazard functions may have big jumps in the first derivative. However, the HARE model for the transformed data typically has fewer knots in time, while the remaining jumps in the first derivative of the baseline hazard function tend to be smaller. Here, as well as in KST, HEFT (see above) is used to get a smooth estimate of the unconditional hazard function.

The HARE analysis of the transformed breast cancer data is summarized in Table 2. Note that $\hat{q}_0 = -\log(1 - \hat{F}_0)$ in this table, where $\hat{F}_0$ is the distribution function corresponding to the estimate of the unconditional hazard function obtained by HEFT. The fitted HARE model has knots located at the transformed times 0.194 and 0.514, which correspond to the real times 1.80 years and 4.68 years, respectively. Note that the model in Table 2 is not a proportional hazards model because of the presence of the basis functions $((0.514 - \hat{q}_0(t))_+ \times \text{ER})$ and $((0.194 - \hat{q}_0(t))_+ \times \text{size})$. Gray[37] noted nonproportionality with respect to ER using time-varying coefficients. In his analysis, he felt that a proportional hazards model with respect to size was appropriate. We investigate this further below. In Table 2 we notice a nonlinear effect (knot) in age

**Table 2** HARE analysis of the transformed breast cancer data

| Basis function | Coefficient | Standard error |
|---|---|---|
| 1 | −0.0443 | 0.3990 |
| ER | 0.426 | 0.119 |
| log(nodes) | 0.686 | 0.070 |
| size | 0.158 | 0.035 |
| age | −0.0401 | 0.0093 |
| $(\text{age}-43)_+$ | 0.0408 | 0.0115 |
| menopause | 0.409 | 0.105 |
| $(0.194-\hat{q}_0(t))_+$ | −6.58 | 1.33 |
| $(0.514-\hat{q}_0(t))_+$ | 2.66 | 0.41 |
| log(nodes) × size | −0.0650 | 0.0181 |
| $(0.514-\hat{q}_0(t))_+ \times$ ER | −2.91 | 0.39 |
| $(0.194-\hat{q}_0(t))_+ \times$ size | 0.878 | 0.266 |

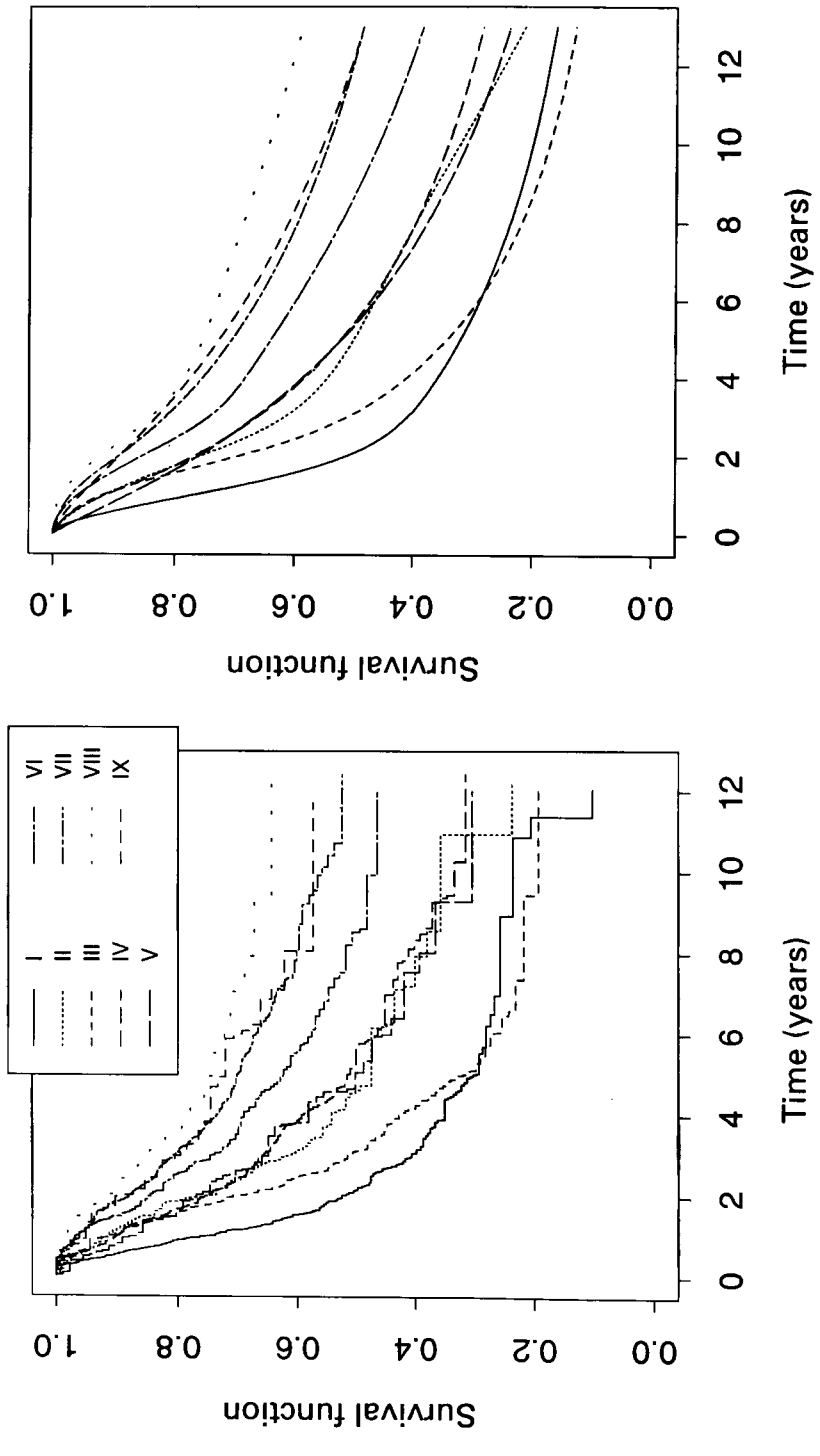


**Figure 2** Estimates of the survival functions for the subjects in each of the terminal nodes for the breast cancer data: left = Kaplan–Meier; right = HEFT.

and an interaction between log(nodes) and size, which were also found in Gray's analysis.

Two of the advantages of HARE are direct results of its user-friendly interface. First, it is easy to graph various functions that may be of interest, such as conditional hazard functions for given sets of covariates. In Figure 3 we show the fitted conditional hazard and survival functions for three sets of covariates. As can be seen from the left side of this figure, ER has a distinctly nonproportional effect: the solid and dotted curves actually cross each other at $t = 3.22$ years. On the other hand, the effect of the number of nodes is proportional.

Secondly, it is very convenient to fit and compare linear proportional hazards models, additive proportional hazards models, proportional hazards models with time-varying coefficients and nonparametric proportional hazards models. As mentioned above, in the analysis of Gray[37] the effect of size was modelled proportionally, while in the HARE analysis an interaction between time and size ended up in the model. To investigate this further, we applied the HARE algorithm forcing an additive model for the log-hazard function. The proportional hazards model that was obtained is partly summarized in Table 3. As can be seen, there is a difference of 54.96 in the BIC value between this model and the full model in Table 2. This difference is substantial, as is the corresponding difference in log-likelihood.

It is also possible to force HARE to fit certain interactions. In particular, we applied the HARE algorithm twice more. Once we required the model to be a proportional hazards model but allowed interactions between covariates, and once we ran the algorithm allowing interactions between covariates and time. These HARE models are also summarized in Table 3. As can be seen from the difference 52.92 in BIC values between the proportional hazards model and the model from Table 2, a proportional hazards model is unsatisfactory. From Table 3 we also note that the model that allows ER $\times$ time interactions but no other interactions with time is not as good as the model in Table 2, but much better than the other models in Table 3. This suggests that the size $\times$ time interaction in the model improves the fit, but it is not as important as the other interactions.

Actually, it turned out that the three restricted HARE models were exactly nested in the full HARE model. This is quite accidental, since the HARE runs are separate and knots are placed independently. When standard methods of statistical inference are applied to adaptive procedures, the results should be considered only as indicative. Nevertheless, if the three restricted models were tested against the unrestricted model, relying on standard $\chi^2$ statistics, then all three smaller models would be strongly rejected in favour of the larger model.

*Proportional hazards analysis*

We also fit a Cox proportional hazards model[6] to the breast cancer data, using a backwards stepwise algorithm. There were two initial sets of covariates that we considered:

1) the regular six covariates (using a log-transform for nodes);
2) the regular six covariates as well as an interaction between log(nodes) and size, suggested by HARE and the survival trees, and an interaction between oestrogen receptor status and menopausal status, suggested by the survival trees.

The results are summarized in Table 4.

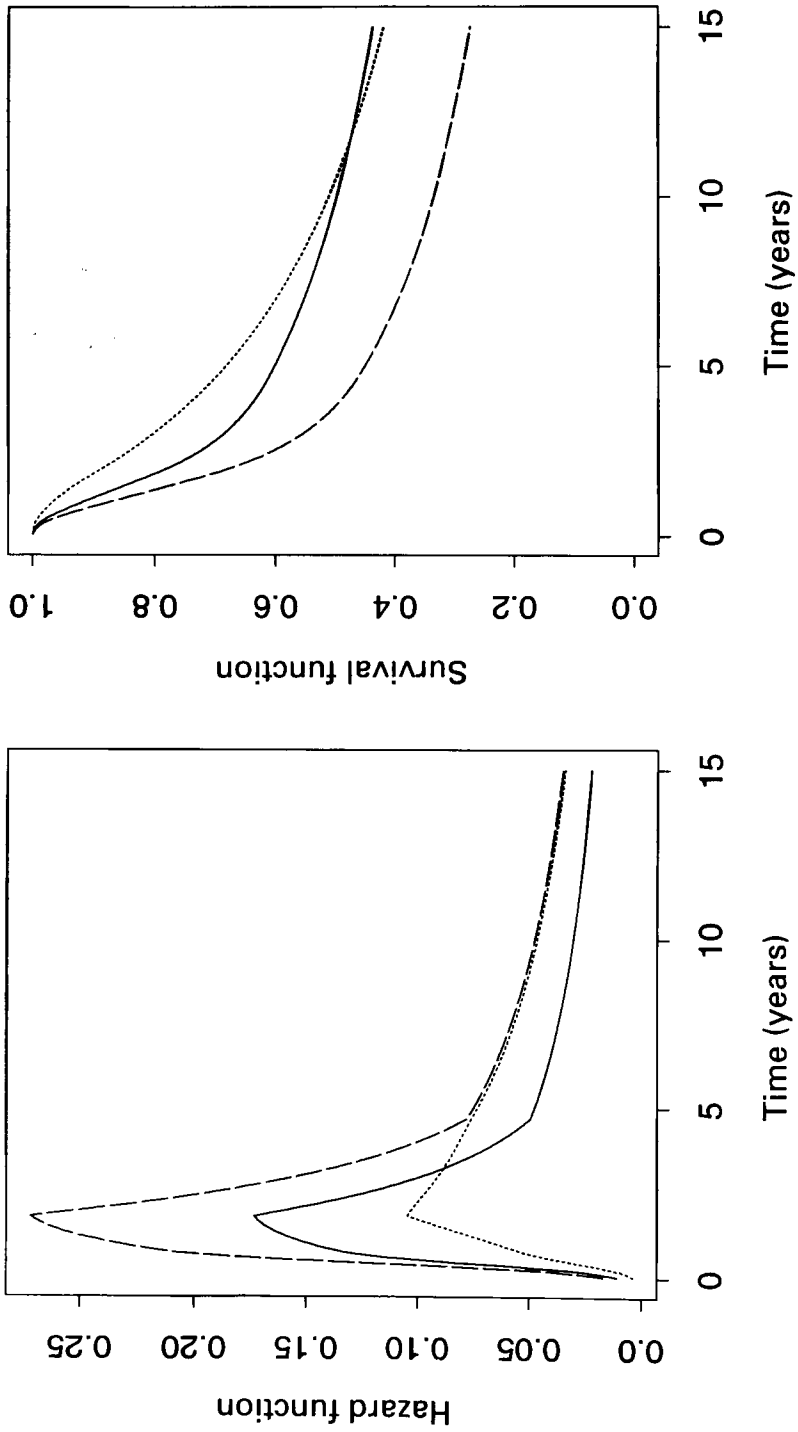


**Figure 3** Fitted hazard and survival functions for a premenopausal woman of age 50 with body mass index 25 and tumour size 3 cm. Solid = 4 nodes and ER negative; dotted = 4 nodes and ER positive; dashed = 10 nodes and ER negative.

**Table 3** Summary of several HARE models for the transformed breast cancer data

| Model | Full | Additive | Prop. Haz. | PH and ER × time |
|---|---|---|---|---|
| dimension | 12 | 7 | 8 | 10 |
| BIC-value | 1906.67 | 1961.73 | 1959.69 | 1911.93 |
| log-likelihood | −906.67 | −953.62 | −943.51 | −917.04 |

*Comparison*

The HARE and survival tree analyses provide us with complementary information. The HARE analysis provides us with a rather complicated, nonproportional model for the conditional hazard function. The survival tree analysis, on the other hand, provides us with a clear indication of the most important variables. It also provides us with a partition of all cases into groups with similar survival characteristics. Whether data is adequately modelled by a proportional hazards model is not addressed in the survival tree method.

For the present example, HARE and survival trees partially confirm each others results. Both methods identify the number of nodes as the most important variable: the first split using survival trees is based on nodes, while nodes has the smallest standard error, relative to its coefficient, in HARE. Both methods find a nodes × size interaction and that the oestrogen receptor status is one of the most useful variables.

The proportional hazards model, on the other hand, misses important aspects of the data. It does not find the nonlinear effect of age on the hazard function. Unless explicitly told to include it, it would not find the strong interaction between nodes and size. Moreover, as shown in the HARE analysis, a proportional hazards model does not fit the data! Interactions between time and oestrogen receptor status and between time and size should be included in a model.

### 4.2 Coronary heart disease data

Our second example has a much higher percentage of censoring than the first example. The data come from a study of mortality due to coronary heart disease (CHD) known as the Western Collaborative Group Study. Of the 3155 men who entered the study 415 died of CHD, 329 died of cancer, 290 died of other causes and 2121 were still alive after 27 years. Since we are only concerned with CHD here, 2740 cases (86.8%) are censored. The study is described in Rosenman *et al.*,[53] Rosenman *et al.*[54] and Ragland and Brand.[55] A tree-based survival analysis, very much along the lines of the one we describe below, has been reported in Carmelli *et al.*[8]

We used the same eight covariates as Carmelli *et al.*[8]: the age at entry into the study, the systolic blood pressure (SBP) at entry into the study, serum cholesterol at entry into the study, a hostility index (originally 1–5, standardized to have mean zero and

**Table 4** Summary of two stepwise proportional hazards models for the breast cancer data

| | Without interactions | | With interactions | |
|---|---|---|---|---|
| Covariate | Coefficient | Standard error | Coefficient | Standard error |
| ER | −0.038 | 0.062 | −0.296 | 0.063 |
| log(nodes) | 0.466 | 0.034 | 0.659 | 0.070 |
| size | 0.087 | 0.017 | 0.183 | 0.035 |
| age | dropped out | | −0.0104 | 0.0043 |
| menopause | 0.285 | 0.062 | 0.481 | 0.100 |
| BMI | dropped out | | dropped out | |
| log(nodes) × size | | | −0.0565 | 0.00181 |
| ER × menopause | | | dropped out | |

standard deviation one), an indicator for behaviour type A (1) or B (0), an indicator for having ever smoked (1 = yes), the body mass index (BMI) and the waist-to-calf ratio (WCR). There were 277 cases (55 being uncensored) for which either BMI or WCR was missing. The survival tree method deals elegantly with missing values using surrogate splits. HARE presently does not have the ability to deal directly with missing values. (It is envisioned that future versions of HARE will have such an ability.)

As in the previous example, we present the survival tree analysis followed by the HARE analysis and a proportional hazards analysis and we conclude the example with a comparison between the three methods.

*Survival tree analysis*

Carmelli *et al.*[8] presented their survival tree analysis of the CHD data. They used logrank statistics to define the splitting rule, while the pruning and tree selection were as in Gordon and Olshen,[13] in which the risk at a node is defined by the fourth power Wasserstein distance between the Kaplan–Meier survival curve of the subjects in a node and a piecewise exponential model with one knot.

An analysis using Intrator's method[18] validated Carmelli's results, using various rank tests ($\rho$ and $\gamma$ combinations) for splitting and the pruning algorithm presented above, which is different from the pruning algorithm used by Carmelli *et al.* Other splits were also examined, specifically those corresponding to the additional variable WCR provided to us and splits determined by quartiles. As in the breast cancer data, we ran the trees several times to determine structure stability.

Most analyses confirmed Carmelli's tree. Other competing structures emerged from the interchange of the first two splits, age (at a 48-year cutpoint) and systolic blood pressure (at a cutpoint of 150 mg/dl). Carmelli *et al.* favoured the root node split of blood pressure. If the root node split was on age, both second level splits were on SBP. For the cases younger than 48 years there were further splits beyond the second (SBP) split. In Figure 4 we summarize Carmelli's tree. Note that the numbers of alive and dead subjects in each node are different from those shown in Figure 1 of Carmelli *et al.*,[8] the reason being that the data we analysed here is slightly different from that analysed by Carmelli *et al.*

In Figure 5 we show estimates for the survival functions for the six terminal nodes. We show both the usual Kaplan–Meier curves (left side) as well as estimates using HEFT, which was described in Section 4.1. As can be seen from Figure 5, the six groups in which the survival tree method divided the data have all distinctly different survival curves. (The groups correspond to terminal nodes I and II may be exceptional. However, those groups were separated on an earlier split of the tree.) We feel that, since the HEFT estimates are nice and smooth, this difference between survival curves is more easily recognized from the HEFT estimates than from the Kaplan–Meier estimates.

Trees from the methods of LeBlanc and Crowley,[22] and of Davis and Anderson[15] were obtained. Leblanc and Crowley's method produced a tree in which the first split is on age, followed by several splits (in both branches) on SBP. After four splits we have the following five nodes:

1) a decision node for which (age $\leq$ 48) and (SBP $\leq$ 151). This node is then split on cholesterol and SBP, yielding terminal nodes similar to nodes IV, V and VI in Figure 4;

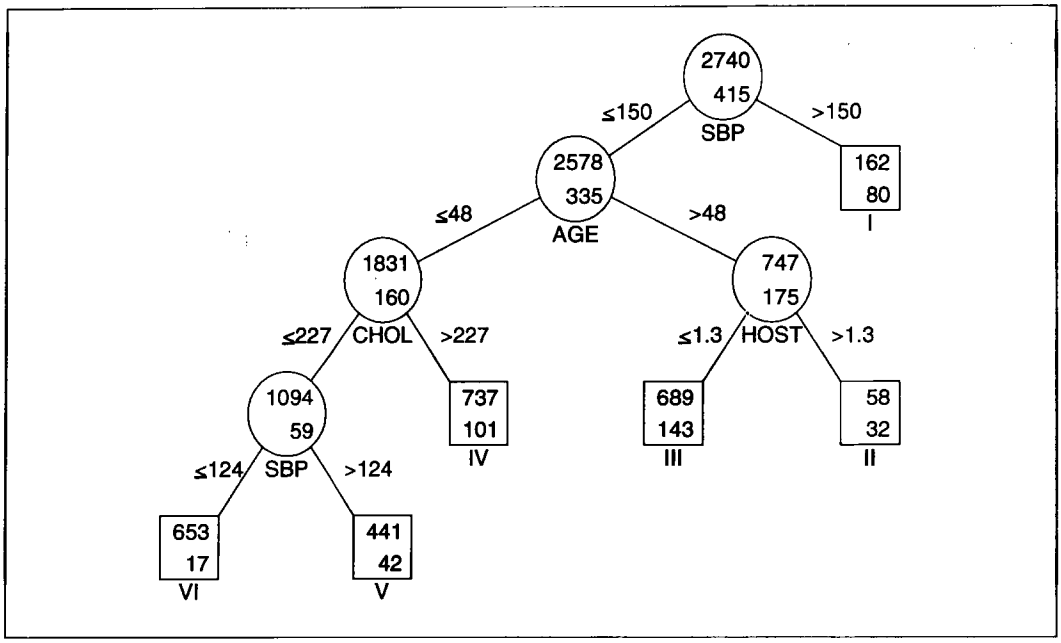


**Figure 4** Survival tree for the coronary heart disease data. Terminal nodes are indicated by squares and nonterminal nodes by circles. Numbers in circles and squares indicate the number of alive (top) and deceased (bottom) cases in that node.

2) a terminal node for which (age $\leq$ 48) and (SBP $>$ 151);
3) a decision node for which (age $>$ 48) and (SBP $\leq$ 133). This node is then split on hostility, yielding terminal nodes similar to nodes II and III in Figure 4;
4) a decision node for which (age $>$ 48) and (133 $<$ SBP $\leq$ 161) which is then split on smoking;
5) a terminal node for which (age $>$ 48) and (SBP $>$ 161).

Thus, as for the breast cancer example, this tree is very similar to the one obtained by Intrator's method.

Davis and Anderson's method produced a tree with a root node split on age (breakpoint at age 48), followed by a split on blood pressure (breakpoint = 151) for older people, and a split on cholesterol (breakpoint = 226) thereafter on blood pressure for the younger people. In total the derived tree had very similar terminal regions as those presented in Figure 4 (hostility was replaced by cholesterol level), although the splitting scheme was somewhat different.

*HARE analysis*

As in the breast cancer example (Section 4.1), we first estimated the unconditional hazard function using HEFT. The estimates of the unconditional hazard rate for all 3155 cases and for the 2878 cases without missing values are shown in the left side of Figure 6. As can be seen from this figure, these estimates are very similar. Note that while we show the conditional hazard rate until 32 years, there are no uncensored observations beyond 28 years. HEFT extrapolates the log-hazard function smoothly. On the right side of Figure 6 we show the corresponding complete unconditional densities that HEFT fit to the data. They are noticeably different, but since we are

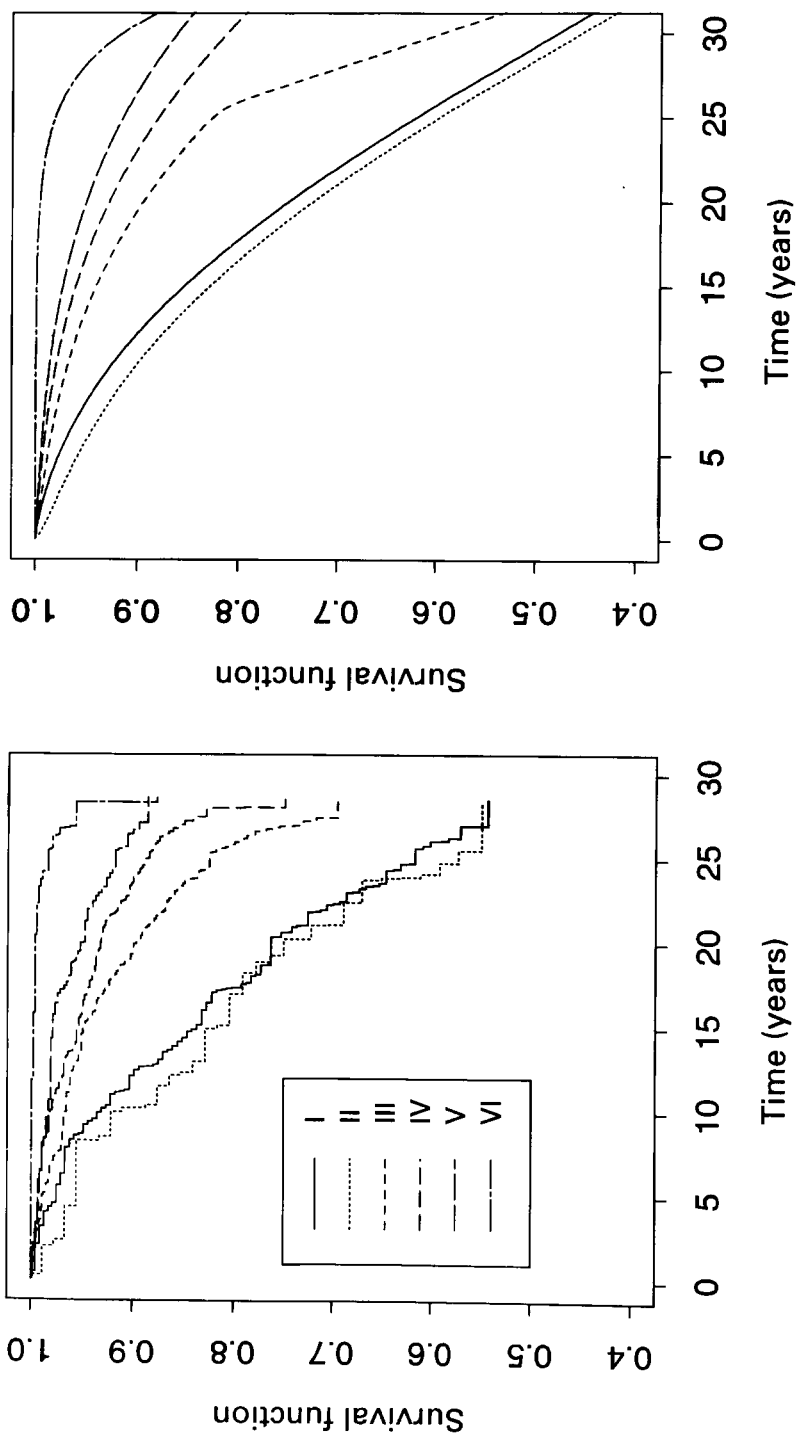


**Figure 5** Estimates of the survival functions for the subjects in each of the terminal nodes for the coronary heart disease data: **left** = Kaplan–Meier; right = HEFT.

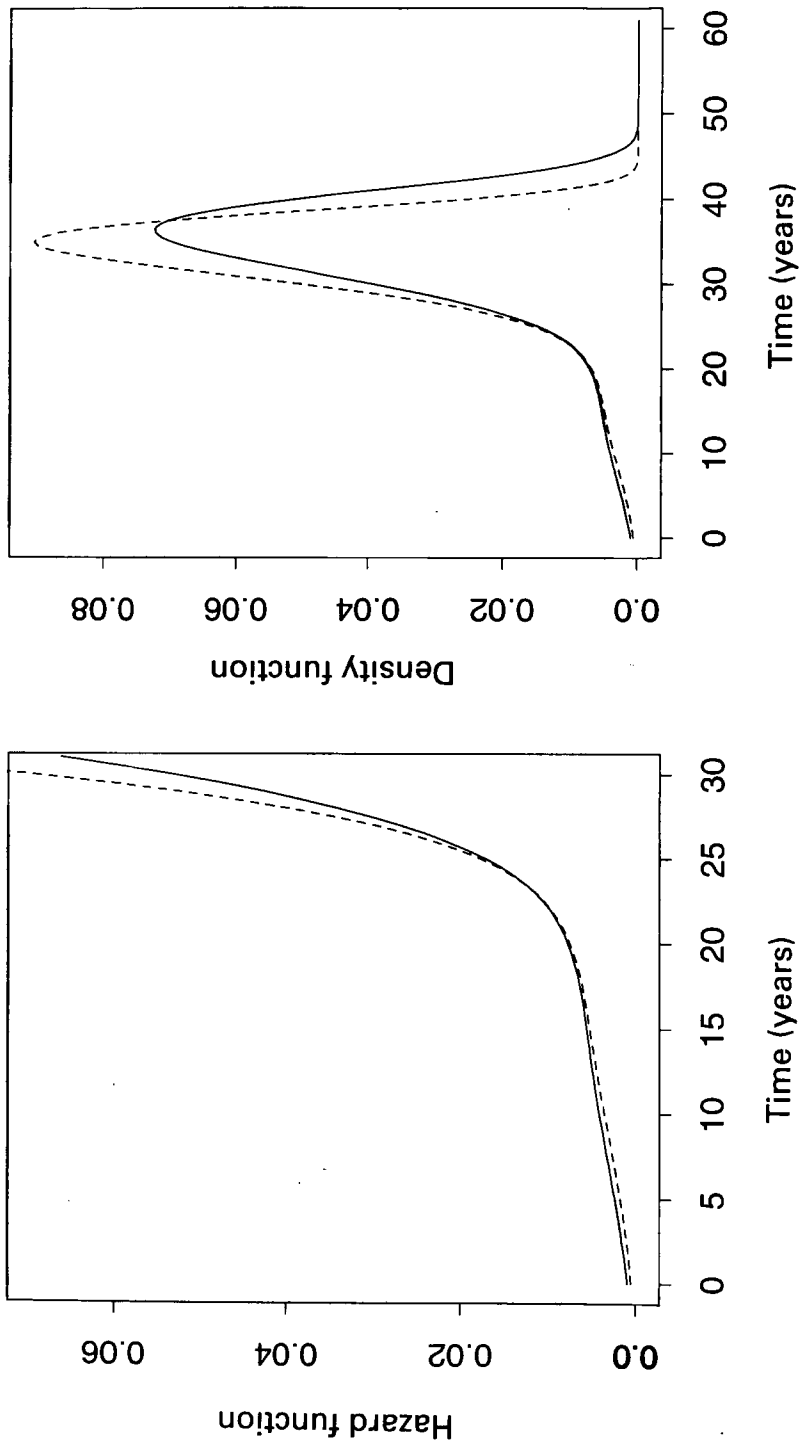


**Figure 6** Estimates of the unconditional hazard and density functions for the coronary heart disease data. Missing values: solid = all cases used; dashed = cases with missing covariates deleted.

**Table 5** HARE analysis of the transformed coronary heart data

| Covariate | Cases with missing values deleted | | All cases, two covariates deleted | |
|---|---|---|---|---|
| | Coefficient | Standard error | Coefficient | Standard error |
| 1 | −9.84 | 0.60 | −9.87 | 0.56 |
| $(0.0058-\hat{q}_0(t))_+$ | −347 | 137 | n/a | n/a |
| SBP | 0.0267 | 0.0029 | 0.0269 | 0.0027 |
| age | 0.0874 | 0.0093 | 0.0888 | 0.0087 |
| serum cholesterol | 0.00814 | 0.00100 | 0.00774 | 0.00095 |
| smoking | 0.432 | 0.144 | 0.436 | 0.135 |

mainly concerned with time less than 30 years, this will not affect our conclusions. In both HEFT estimates about 81% of the probability mass is beyond 28 years.

Since the Western Collaborative Group Study is ongoing, the censoring is essentially Type I.[56] In particular, we note that of the 2121 men who have not died, 2091 (98.6%) have been in the study for over 20 years, while of the 415 men who died of CHD 229 (55.2%) died before they had been in the study for 20 years. This type of censoring substantially limits the conclusions that we can reach: our effective sample size is much smaller than 3155 and there is no data beyond the 20th percentile of the unconditional distribution function.

Keeping this in mind, we applied HARE to the transformed data, with the missing cases deleted. The HARE model is summarized in Table 5. As it turned out, none of the covariates SBP, age, serum cholesterol and smoking that ended up in the model have missing values. This allows us to apply HARE to the complete dataset using just the six covariates that have no missing values. This model is also summarized in Table 5. As can be seen, both models are linear proportional hazards models and they are nearly identical to each other, the only difference being one extra basis function when the cases with missing values are deleted. Since this basis function does not depend on a covariate, it effectively corresponds to a knot in the baseline hazard function. Note that when there are no knots in time and no interactions between time and covariates, as is the case for the model based on all 3155 cases, the estimate of the unconditional hazard function shown here in Figure 6 is, except for a scaling, a smooth estimate of the baseline hazard function in the proportional hazards model.

In Kooperberg,[46] where HARE was applied to some datasets with large amounts of interval censoring, linear proportional hazards models were obtained. There it was hypothesized that this was because of the 'lack of signal' due to the large amount of censoring. It also seems possible that HARE would not find certain interactions because it never entered them in the model since other functions were considered more useful. In Carmelli *et al.*[8] the survival tree ended up including two three-way interactions: one between SBP, age and serum cholesterol and one between SBP, age and hostility. To compare a model with (some of) these interactions with the HARE model, we fit a HARE model forcing it to look as much as possible like the model in Carmelli *et al.*[8]

Specifically, we included only the covariates SBP, age, serum cholesterol and hostility, we forced the model to be a proportional hazards model, and we did not allow an interaction between serum cholesterol and hostility. (Since all missing data were in BMI and WCR, we included all cases in the analysis.) As it turned out, only the constant and the covariates SBP, age and serum cholesterol, all linear, ended up in the HARE model. If the penalty term $a$ for AIC in equation (3.2) were reduced from

$\log n = \log 3155 \approx 8.06$ (BIC) to below $5.42 \approx \log 226$, hostility would also be included in the model. This correspond to a $p$ value of about 0.02. Although we usually would include a variable with such a $p$ value, we are more cautious if our procedure is as adaptive as the HARE procedure, since we consider many candidates for basis functions. Furthermore, as mentioned above, the $p$ value of hostility becomes 0.04 once smoking has been entered in the model.

If the penalty term in equation (3.2) were reduced to below $4.92 \approx \log 137$ the interaction between SBP and serum cholesterol would be included in the selected HARE model. This corresponds to a $p$ value of about 0.026. No other interaction had a $p$ value below 0.1. In summary, except for some support for a SBP $\times$ serum cholesterol interaction, the HARE analysis does not support the interactions found using survival trees.

*Proportional hazards analysis*

If a proportional hazards model is fit to the coronary heart disease data, the four variables SBP, age, serum cholesterol and smoking in the model are found to be very significant (hostility has a $p$ value of about 0.04). If a linear proportional hazards model with just these four variables is fit, the coefficients are virtually the same as in Table 5. It is interesting to note that in the survival tree approach[8] smoking did not end up in the model.

*Comparison*

Whereas for our first example the main results from the survival tree analysis were confirmed by the HARE analysis, this is much less the case for the second example. Both analyses recognize that SBP, age and serum cholesterol are the most important variables, but the survival tree method comes up with two three-factor interactions, while the HARE analysis uses an additive model and does not confirm the interactions. Also, HARE uses smoking as the fourth most important variable, while the survival trees use hostility. HARE and the proportional hazards model are in complete agreement about this dataset.

We started this example by noting that it has a particularly high percentage for censoring. This may be one reason for the discrepancy between HARE and the proportional hazards analysis on one side and the survival trees on the other side.

## 5 Discussion

When analysing real data, usually model assumptions are hard to verify. Thus, it is important to analyse data by several methods. We believe that both survival trees and hazard regression are worthwhile tools for the investigator to use in order to gain new insights about the data that may easily be missed by immediately applying standard proportional hazards models, as we saw for the breast cancer data.

The basic ingredients that are involved in actual applications using these methods have been introduced here. Details about the programs can be found in the papers where the methods were introduced. HARE and HEFT are available from statlib (`statlib@stat.cmu.edu`). Intrator's program[18] is available by sending email to `msorna@olive.huji.ac.il`. Davis' program[15] is available by sending email directly to `rdaids@sdac.harvard.edu`, LeBlanc and Crowley's program[22] is available by sending email to `mikel@orca.fhcrc.org`, Segal's program[19] is avail-

able by sending email to `mark@segal.ucsf.edu`, Zhang's program[17] is available by sending email to `heping@peace.med.yale,edu`, and Gordon and Olshen's program[13] is available solely for academic use by sending email to `dstein@saturn.sdsu.edu`.

Hazard regression and survival trees are two new methods for the analysis of survival data that deserve a place in the toolbox of the survival analyst. These methods have their specific strengths. In particular, one of the appealing features of HARE is that it provides an automatic check for the appropriateness of a proportional hazards model. When a proportional hazards model is appropriate, HARE, especially when used in conjunction with HEFT, provides a smooth estimate for the underlying baseline hazard function. It also provides a MARS-like[2] model for the conditional hazards function. Another strength is a graphical interface that makes it very easy to look at curves such as conditional hazards functions. Thus a HARE model is potentially useful for a health care practitioner in coming up with a prognosis for a particular patient.

Both HARE and survival tree methods reveal the hidden structure of the data by reducing the number of important predictors. HARE does this by fitting a model with main effect and two-factor interaction terms. On the other hand, survival trees provide a hierarchical set of questions about the population associating it with a specific survival distribution, without assuming any parametric form of regression dependence. The important characteristics of terminal nodes are the estimated survival probabilities, the probability of survival past a certain point, and quantiles of the survival distribution.

An advantage of the survival tree method is that it has a number of features that the present implementation of HARE lacks because survival trees are based on the more established CART methodology. Examples are ranking of the covariates affecting the process in order of importance (the importance ranking, though, is not a complete answer since variables tend to act in concert and not alone) and an elegant way to deal with missing values based on surrogate splits. However, in a proposed commercial implementation HARE should be able to deal with missing values in a way that is similar to the function `no.gam.replace()` in S,[50] and it will have an importance measure for the covariates based on an ANOVA decomposition.[30]

There is a similarity between the procedure by which variables are selected through the splitting and pruning algorithm of survival trees and the stepwise addition and deletion algorithm of HARE. But there are also major differences between survival trees and HARE. While most variables selected by HARE act on the whole data, the variables in a survival tree act on subsets of the data. In general, a method that looks for effects within subsets may have a greater ability to detect interaction effects among the variables than do methods in which variables act over the entire range of the vector of covariates. Interactions that have an effect only within a subset may be important in identifying (smaller) subsets with unique survival that would otherwise go unnoticed. On the other hand, HARE allows variables to act additively and linearly, a capability that the survival tree method lacks.

**Acknowledgements**

Charles Kooperberg was supported in part by the NSF under the grant DMS-94.03771. Orna Intrator was partially supported by grant 359/93-1 from the Israeli Academy of Science. The data for the example in Section 4.1 was kindly provided by

Robert Gray of the Eastern Cooperative Oncology Group. The data for the example in Section 4.2 was kindly provided by Dorit Carmelli of the Western Collaborative Group Study. We thank Michael LeBlanc, Richard A Olshen, Charles J Stone and Heping Zhang for their help in the preparation of this review.

## References

1 Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and regression trees.* Pacific Grove, CA: Wadsworth, 1984.
2 Friedman JH. Multivariate regression splines (with discussion). *The Annals of Statistics* 1988; **19**: 1–141.
3 Kooperberg C, Stone CJ. Logspline density estimation for censored data. *Journal of Computational and Graphical Statistics* 1992; **1**: 301–28.
4 Kooperberg C, Stone CJ, Truong YK. Logspline estimation of a possibly mixed spectral distribution. *Journal of Time Series Analysis* 1995; **16**: 359–88.
5 Bose S, Kooperberg C, Stone CJ. Polychotomous regression. University of Washington, Department of Statistics, Technical Report, No. 288, 1995.
6 Cox DR. Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B* 1972; **34**: 187–220.
7 Kwak, LW, Halpern J, Olshen RA, Horning SJ. Prognostic significance of actual dose intensity in diffuse large cell lymphoma: Results of tree-structured survival analysis. *Journal of Clinical Oncology* 1990; **8**: 963–77.
8 Carmelli D, Halpern J, Swan GE *et al.* 27-Year mortality in the Western Collaborative Group Study: construction of risk groups by recursive partitioning. *Journal of Clinical Epidemiology* 1991; **44**: 1341–51.
9 Piette JD, Intrator O, Zierler S, Mor V, Stein M. Differences in case fatality rates for Aids patients: application of a new methodology for survival research. *Epidemiology* 1992; **3**: 310–18.
10 Mor V, Intrator O, Laliberte LL. Factors affecting conversion rates to Medicaid among new admissions to nursing homes. *Health Services Research* 1993; **28**: 1–25.
11 Schumacher M, Schmoor C, Sauerbrei W *et al.* The prognostic effect of histological tumor grade in node negative breast cancer patients. *Breast Cancer Research and Treatment* 1993; **25**: 235–45.
12 Curran WJ, Scott CB, Horton J *et al.* Recursive partitioning analysis of prognostic factors in three radiation therapy oncology group malignant glioma trials. *Journal of the National Cancer Institute* 1993; **85**(9): 704–10.
13 Gordon L, Olshen RA. Tree-structured survival analysis. *Cancer Treatment Reports* 1985; **69**: 1065–69.
14 Ciampi A, Chang, CH, Hogg S, McKinney S. Recursive partitioning: a versatile method for exploratory data analysis in Biostatistics. In: MacNeil IB, Umphrey G eds. *Proceedings from Joshi Festschrift.* Amsterdam: North Holland, 1987: 23–50.
15 Davis RB, Anderson JR. Exponential survival trees. *Statistics in Medicine* 1989; **8**: 947–61.
16 LeBlanc M, Crowley J. Relative risk trees for censored data. *Biometrics* 1992; **48**: 411–25.
17 Zhang HP. Splitting criteria in survival trees. In: *10th Workshop on Statistical Modeling.* Lecture Notes in Statistics Series. New York: Springer-Verlag, 1995: 305–13.
18 Intrator O. Exploratory trees for semi-markov processes. *Proceedings of the 23rd symposium on the interface of computing science and statistics* 1991; 352–55.
19 Segal MR. Regression trees for censored data. *Biometrics* 1988; **44**: 35–47.
20 Segal MR, Bloch DA. A comparison of estimated proportional hazards models and regression trees. *Statistics in Medicine* 1989; **8**: 539–50.
21 Butler JH, Gilpin E, Gordon L, Olshen RA. Tree structured survival analysis, II. *Technical Report, Stanford University,* 1991.
22 LeBlanc M, Crowley J. A review of tree-based prognostic models. In: *New Advances in the Design and Analysis of Clinical Trials Data.* Kluwer Academic Publishers, 1995: 113–24.
23 Ciampi A, Thiffault J, Nakache J-P, Asselain B. Stratification by stepwise regression, correspondence analysis and recursive partition. *Computational Statistics and Data Analysis* 1986; **4**: 185–204.
24 Harrington DP, Fleming TR. A class of rank test procedures for censored survival data. *Biometrika* 1982; **69**: 553–66.
25 Fleming TR, Augustine GA, Elcombe SA, Offord KP. The SURVDIF procedure. *SAS/ SUGI Users manual, version 5,* 1986.

26 Fleming TR, Harrington DP, O'Sullivan M. Supremum versions of the log-rank and generalized Wilcoxon statistics. *Journal of the American Statistical Association* 1987; **82**: 312–20.

27 Breslow N. Contribution to the discussion of a paper by DR Cox. *Journal of the Royal Statistical Society, Series B* 1972; **34**: 216–17.

28 Nelson, W. On estimating the distribution of a random vector when only the coordinate is observable. *Technometrics* 1969; **12**: 923–24.

29 Eubank RL. *Spline smoothing and nonparametric regression.* New York: Marcel Dekker, 1988.

30 Hastie TJ, Tibshirani R. *Generalized additive models.* London: Chapman and Hall, 1990.

31 Wahba G. *Spline models for observational data.* Philadelphia: SIAM, 1990.

32 Anderson JA, Senthilselvan A. Smooth estimates for the hazard function. *Journal of the Royal Statistical Society, Series B* 1980; **42**: 322–27.

33 Whittemore AS, Keller JB. Survival estimation using splines. *Biometrics* 1986; **42**: 495–506.

34 Senthilselvan A. Penalized likelihood estimation of hazard and intensity functions. *Journal of the Royal Statistical Society, Series B* 1987; **49**: 170–74.

35 O'Sullivan F. Nonparametric estimation of relative risk using splines and cross-validation. *SIAM Journal of Scientific and Statistical Computing* 1988; **9**: 531–42.

36 O'Sullivan F. Fast computation of fully automated log-density and log-hazard estimators. *SIAM Journal of Scientific and Statistical Computing* 1988; **9**: 363–79.

37 Gray RJ. Flexible methods for analyzing survival data using splines, with applications to breast cancer prognosis. *Journal of the American Statistical Association* 1992; **87**: 942–51.

38 Hastie TJ, Tibshirani R. Varying-coefficient models (with discussion). *Journal of the Royal Statistical Society, Series B* 1993; **55**: 757–800.

39 Gu C. Penalized likelihood hazard estimation. Technical Report No. 91-58, Department of Statistics, Purdue University, 1991.

40 Gu C. Structural multivariate function estimation: some automatic density and hazard estimates. Purdue University, Department of Statistics, Technical Report, 1994.

41 Smith PL. Curve fitting and modeling with splines using statistical variable selection methods. NASA Report 166034, Langley Research Center, Hampla, VA: NASA, 1982.

42 Etezadi-Amoli J, Ciampi A. Extended hazard regression for censored survival data with covariates: a spline approximation for the baseline hazard function. *Biometrics* 1987; **43**: 181–92.

43 Efron B. Logistic regression, survival analysis and the Kaplan–Meier curve. *Journal of American Statistical Association* 1988; **83**: 414–25.

44 Abrahamowicz M, Ciampi A, Ramsay JO. Nonparametric density estimation for censored survival data: Regression-spline approach. *The Canadian Journal of Statistics* 1991; **20**: 171–85.

45 Kooperberg C, Stone CJ, Truong YK. Hazard regression. *Journal of the American Statistical Association* 1995; **90**: 78–94.

46 Kooperberg C. Hazard regression for interval censored data. In: *The Proceedings of the International Biometrics Conference*, Hamilton, Ontario, 1994: 81–96.

47 Kooperberg C, Stone CJ, Truong YK. The $L_2$ rate of convergence of hazard regression. *Scandinavian Journal of Statistics* 1995; **22**: 143–57.

48 Miller RG. *Survival analysis.* New York: Wiley, 1981.

49 Rao CR. *Linear statistical inference and its applications,* second edition. New York: Wiley, 1973.

50 Chambers JM, Hastie TJ. *Statistical models in S.* Pacific Grove, CA: Wadsworth, 1992.

51 Becker RA, Chambers JM, Wilks AR. *The new S language.* Pacific Grove, CA: Wadsworth, 1988.

52 Lindley DV. Contribution to the discussion of Regression models and life tables. *Journal of the Royal Statistical Society, Series B* 1972; **34**: 208–209.

53 Rosenman RH, Friedman M, Strauss R *et al.* A predictive study of coronary heart disease: The Western Collaborative Group Study. *Journal of the American Medical Association* 1964; **189**: 15–22.

54 Rosenman RH, Brand RJ, Sholtz RI, Friedman M. Multivariate prediction of coronary heart disease during 8.5 year follow-up in the Western Collaborative Group Study. *American Journal of Cardiology* 1976; **37**: 903–10.

55 Ragland DR, Brand RJ. Coronary heart disease mortality in the Western Collaborative Group Study: follow-up experience of 22 years. *American Journal of Epidemiology* 1988; **127**: 462–75.

56 Cox DR, Oakes D. *Survival analysis.* London: Chapman and Hall, 1983.