1

Decision Tree for Prognostic Classification of Multivariate Survival Data and Competing Risks

N. A. Ibrahim[1] and A. Kudus[2]
[1]*Institute for Mathematical Research and Department of Mathematics, Faculty of Science, Universiti Putra Malaysia, Serdang*
[2]*Institute for Mathematical Research, Universiti Putra Malaysia and Department of Statistics, Bandung Islamic University, Bandung*
[1]*Malaysia*
[2]*Indonesia*

1. Introduction

Decision tree (DT) is one way to represent rules underlying data. It is the most popular tool for exploring complex data structures. Besides that it has become one of the most flexible, intuitive and powerful data analytic tools for determining distinct prognostic subgroups with similar outcome within each subgroup but different outcomes between the subgroups (i.e., prognostic grouping of patients). It is hierarchical, sequential classification structures that recursively partition the set of observations. Prognostic groups are important in assessing disease heterogeneity and for design and stratification of future clinical trials. Because patterns of medical treatment are changing so rapidly, it is important that the results of the present analysis be applicable to contemporary patients.

Due to their mathematical simplicity, linear regression for continuous data, logistic regression for binary data, proportional hazard regression for censored survival data, marginal and frailty regression for multivariate survival data, and proportional subdistribution hazard regression for competing risks data are among the most commonly used statistical methods. These parametric and semiparametric regression methods, however, may not lead to faithful data descriptions when the underlying assumptions are not satisfied. Sometimes, model interpretation can be problematic in the presence of high-order interactions among predictors.

DT has evolved to relax or remove the restrictive assumptions. In many cases, DT is used to explore data structures and to derive parsimonious models. DT is selected to analyze the data rather than the traditional regression analysis for several reasons. Discovery of interactions is difficult using traditional regression, because the interactions must be specified a priori. In contrast, DT automatically detects important interactions. Furthermore, unlike traditional regression analysis, DT is useful in uncovering variables that may be largely operative within a specific patient subgroup but may have minimal effect or none in other patient subgroups. Also, DT provides a superior means for prognostic classification. Rather than fitting a model to the data, DT sequentially divides the patient group into two

subgroups based on prognostic factor values (e.g., tumor size < 2 cm vs tumor size ≥ 2 cm). The repeated partitioning creates "bins" of observations that are approximately homogeneous. This permits the use of some summary functions (e.g., Kaplan-Meier or cumulative incidence function (CIF)) to compare prognosis between the "bins." The combination of binning and the interpretability of the resulting tree structure make DT extremely well suited for developing prognostic stratifications.

The landmark work of DT in statistical community is the Classification and Regression Trees (CART) methodology of Breiman et al. (1984). A different approach was C4.5 proposed by Quinlan (1993). Original DT method was used in classification and regression for categorical and continuous response variable, respectively. In a clinical setting, however, the outcome of primary interest is often duration of survival, time to event, or some other incomplete (that is, censored) outcome. Therefore, several authors have developed extensions of original DT in the setting of censored survival data (Banerjee & Noone, 2008).

In science and technology, interest often lies in studying processes which generate events repeatedly over time. Such processes are referred to as recurrent event processes and the data they provide are called recurrent event data which includes in multivariate survival data. Such data arise frequently in medical studies, where information is often available on many individuals, each of whom may experience transient clinical events repeatedly over a period of observation. Examples include the occurrence of asthma attacks in respirology trials, epileptic seizures in neurology studies, and fractures in osteoporosis studies. In business, examples include the filing of warranty claims on automobiles, or insurance claims for policy holders. Since multivariate survival times frequently arise when individuals under observation are naturally clustered or when each individual might experience multiple events, then further extensions of DT are developed for such kind of data.

In some studies, patients may be simultaneously exposed to several events, each competing for their mortality or morbidity. For example, suppose that a group of patients diagnosed with heart disease is followed in order to observe a myocardial infarction (MI). If by the end of the study each patient was either observed to have MI or was alive and well, then the usual survival techniques can be applied. In real life, however, some patients may die from other causes before experiencing an MI. This is a competing risks situation because death from other causes prohibits the occurrence of MI. MI is considered the event of interest, while death from other causes is considered a competing risk. The group of patients' dead of other causes cannot be considered censored, since their observations are not incomplete.

The extension of DT can also be employed for competing risks survival time data. These extensions can make one apply the technique to clinical trial data to aid in the development of prognostic classifications for chronic diseases.

This chapter will cover DT for multivariate and competing risks survival time data as well as their application in the development of medical prognosis. Two kinds of multivariate survival time regression model, i.e. marginal and frailty regression model, have their own DT extensions. Whereas, the extension of DT for competing risks has two types of tree. First, the "single event" DT is developed based on splitting function using one event only. Second, the "composite events" tree which use all the events jointly.

2. Decision Tree

A DT is a tree-like structure used for classification, decision theory, clustering, and prediction functions. It depicts rules for dividing data into groups based on the regularities in the data. A DT can be used for categorical and continuous response variables. When the response variables are continuous, the DT is often referred to as a regression tree. If the response variables are categorical, it is called a classification tree. However, the same concepts apply to both types of trees. DTs are widely used in computer science for data structures, in medical sciences for diagnosis, in botany for classification, in psychology for decision theory, and in economic analysis for evaluating investment alternatives.

DTs learn from data and generate models containing explicit rule-like relationships among the variables. DT algorithms begin with the entire set of data, split the data into two or more subsets by testing the value of a predictor variable, and then repeatedly split each subset into finer subsets until the split size reaches an appropriate level. The entire modeling process can be illustrated in a tree-like structure.

A DT model consists of two parts: creating the tree and applying the tree to the data. To achieve this, DTs use several different algorithms. The most popular algorithm in the statistical community is Classification and Regression Trees (CART) (Breiman et al., 1984). This algorithm helps DTs gain credibility and acceptance in the statistics community. It creates binary splits on nominal or interval predictor variables for a nominal, ordinal, or interval response. The most widely-used algorithms by computer scientists are ID3, C4.5, and C5.0 (Quinlan, 1993). The first version of C4.5 and C5.0 were limited to categorical predictors; however, the most recent versions are similar to CART. Other algorithms include Chi-Square Automatic Interaction Detection (CHAID) for categorical response (Kass, 1980), CLS, AID, TREEDISC, Angoss KnowledgeSEEKER, CRUISE, GUIDE and QUEST (Loh, 2008). These algorithms use different approaches for splitting variables. CART, CRUISE, GUIDE and QUEST use the statistical approach, while CLS, ID3, and C4.5 use an approach in which the number of branches off an internal node is equal to the number of possible categories. Another common approach, used by AID, CHAID, and TREEDISC, is the one in which the number of nodes on an internal node varies from two to the maximum number of possible categories. Angoss KnowledgeSEEKER uses a combination of these approaches. Each algorithm employs different mathematical processes to determine how to group and rank variables.

Let us illustrate the DT method in a simplified example of credit evaluation. Suppose a credit card issuer wants to develop a model that can be used for evaluating potential candidates based on its historical customer data. The company's main concern is the default of payment by a cardholder. Therefore, the model should be able to help the company classify a candidate as a possible defaulter or not. The database may contain millions of records and hundreds of fields. A fragment of such a database is shown in Table 1. The input variables include income, age, education, occupation, and many others, determined by some quantitative or qualitative methods. The model building process is illustrated in the tree structure in Figure 1.

The DT algorithm first selects a variable, income, to split the dataset into two subsets. This variable, and also the splitting value of $31,000, is selected by a splitting criterion of the algorithm. There exists many splitting criteria (Mingers, 1989). The basic principle of these criteria is that they all attempt to divide the data into clusters such that variations within each cluster are minimized and variations between the clusters are maximized. The follow

up splits are similar to the first one. The process continues until an appropriate tree size is reached. Figure 1 shows a segment of the DT. Based on this tree model, a candidate with income at least $31,000 and at least college degree is unlikely to default the payment; but a self-employed candidate whose income is less than $31,000 and age is less than 28 is more likely to default.

Name	Age	Income	Education	Occupation	...	Default
Andrew	42	45600	College	Manager	...	No
Allison	26	29000	High School	Self Owned	...	Yes
Sabrina	58	36800	High School	Clerk	...	No
Andy	35	37300	College	Engineer	...	No
…	...	...	...	...	...	...

Table 1. Partial records and fields of a database table for credit evaluation

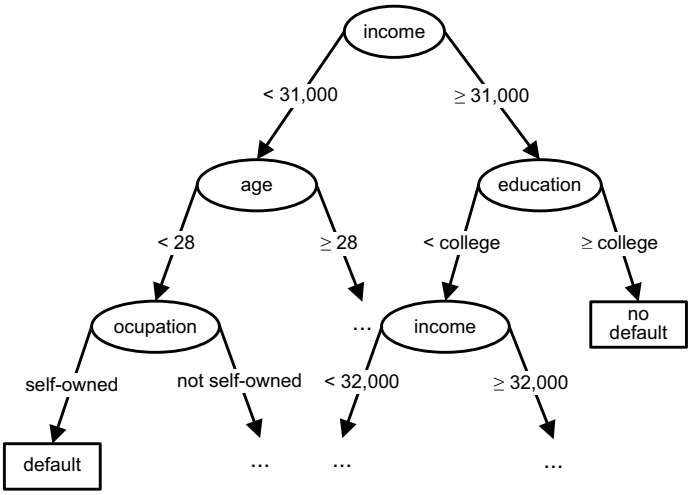

Fig. 1. The decision tree for the credit evaluation example

We begin with a discussion of the general structure of a popular DT algorithm in statistical community, i.e. CART model. A CART model describes the conditional distribution of y given X, where y is the response variable and X is a set of predictor variables ($X = (X_1, X_2, \ldots, Xp)$). This model has two main components: a tree T with b terminal nodes, and a parameter $\Theta = (\theta_1, \theta_2, \ldots, \theta_b) \subset R^k$ which associates the parameter values θ_m, with the m^{th} terminal node. Thus a tree model is fully specified by the pair (T, Θ). If X lies in the region corresponding to the m^{th} terminal node then $y \mid X$ has the distribution $f(y \mid \theta_m)$, where we use f to represent a conditional distribution indexed by θ_m. The model is called a regression tree or a classification tree according to whether the response y is quantitative or qualitative, respectively.

2.1 Splitting a tree

The DT T subdivides the predictor variable space as follows. Each internal node has an associated splitting rule which uses a predictor to assign observations to either its left or

right child node. The internal nodes are thus partitioned into two subsequent nodes using the splitting rule. For quantitative predictors, the splitting rule is based on a split rule c, and assigns observations for which $\{x_i < c\}$ or $\{x_i \geq c\}$ to the left or right child node respectively. For qualitative predictors, the splitting rule is based on a category subset C, and assigns observations for which $\{x_i \in C\}$ or $\{x_i \notin C\}$ to the left or right child node, respectively.
For a regression tree, conventional algorithm models the response in each region R_m as a constant θ_m. Thus the overall tree model can be expressed as (Hastie et al., 2001):

$$f(x) = \sum_{m=1}^{b} \theta_m I(X \in R_m) \tag{1}$$

where R_m, m = 1, 2,…,b consist of a partition of the predictors space, and therefore representing the space of b terminal nodes. If we adopt the method of minimizing the sum of squares $\sum (y_i - f(X_i))^2$ as our criterion to characterize the best split, it is easy to see that the best $\hat{\theta}_m$, is just the average of y_i in region R_m:

$$\hat{\theta}_m = ave(y_i \mid X_i \in R_m) = \frac{1}{N_m} \sum_{X_i \in R_m} y_i \tag{2}$$

where N_m is the number of observations falling in node m. The residual sum of squares is

$$Q_m(T) = \frac{1}{N_m} \sum_{X_i \in R_m} \left(y_i - \hat{\theta}_m\right)^2 \tag{3}$$

which will serve as an impurity measure for regression trees.
If the response is a factor taking outcomes 1,2,...,K, the impurity measure $Q_m(T)$, defined in (3) is not suitable. Instead, we represent a region R_m with N_m observations with

$$\hat{p}_{mk}(T) = \frac{1}{N_m} \sum_{X_i \in R_m} I(y_i = k) \tag{4}$$

which is the proportion of class k($k \in \{1, 2,\ldots,K\}$) observations in node m. We classify the observations in node m to a class $k(m) = \arg\max_k \hat{p}_{mk}$, the majority class in node m. Different measures $Q_m(T)$ of node impurity include the following (Hastie et al., 2001):

$$\begin{aligned} &\text{Misclassification error: } \frac{1}{N_m} \sum_{i \in R_m} I(y_i \notin k) = 1 - \hat{p}_{mk} \\ &\text{Gini index: } \sum_{k \neq k'} \hat{p}_{mk} \hat{p}_{mk'} = \sum_{k=1}^{K} \hat{p}_{mk}(1 - \hat{p}_{mk}) \\ &\text{Cross-entropy or deviance: } \sum_{k=1}^{K} \hat{p}_{mk} \log \hat{p}_{mk} \end{aligned} \tag{5}$$

For binary outcomes, if p is the proportion of the second class, these three measures are $1 - \max(p, 1-p)$, $2p(1-p)$ and $-p \log p - (1-p)\log(1-p)$, respectively.
All three definitions of impurity are concave, having minimums at $p = 0$ and $p = 1$ and a maximum at $p = 0.5$. Entropy and the Gini index are the most common, and generally give very similar results except when there are two response categories.

2.2 Pruning a tree

To be consistent with conventional notations, let's define the impurity of a node h as $I(h)$ ((3) for a regression tree, and any one in (5) for a classification tree). We then choose the split with maximal impurity reduction

$$\Delta I = I(h) - p(h_L)I(h_L) - p(h_R)I(h_R) \tag{6}$$

where h_L and h_R are the left and right children nodes of h and $p(h)$ is proportion of sample that falls in node h.
How large should we grow the tree then? Clearly a very large tree might overfit the data, while a small tree may not be able to capture the important structure. Tree size is a tuning parameter governing the model's complexity, and the optimal tree size should be adaptively chosen from the data. One approach would be to continue the splitting procedures until the decrease on impurity due to the split exceeds some threshold. This strategy is too short-sighted, but however, a seeming worthless split might lead to a very good split below it.
The preferred strategy is to grow a large tree T_0, stopping the splitting process when some minimum number of observations in a terminal node (say 10) is reached. Then this large tree is pruned using pruning algorithm, such as cost-complexity or split complexity pruning algorithm.
To prune large tree T_0 by using cost-complexity algorithm, we define a subtree $T \prec T_0$ to be any tree that can be obtained by pruning T_0, and define $\widetilde{T}$ to be the set of terminal nodes of T. That is, collapsing any number of its terminal nodes. As before, we index terminal nodes by m, with node m representing region R_m. Let $|\widetilde{T}|$ denotes the number of terminal nodes in T ($|\widetilde{T}| = b$). We use $|\widetilde{T}|$ instead of b following the "conventional" notation and define the risk of trees and define cost of tree as

$$\begin{aligned} &\text{Regression tree: } R(T) = \sum_{m=1}^{|\widetilde{T}|} N_m Q_m(T), \\ &\text{Classification tree: } R(T) = \sum_{h \in \widetilde{T}} p(h) r(h), \end{aligned} \tag{7}$$

where $r(h)$ measures the impurity of node h in a classification tree (can be any one in (5)).
We define the cost complexity criterion (Breiman et al., 1984)

$$R_\alpha(T) = R(T) + \alpha |\widetilde{T}| \tag{8}$$

where $\alpha (> 0)$ is the complexity parameter. The idea is, for each α, find the subtree $T_\alpha \prec T_0$ to

minimize $R_\alpha(T)$. The tuning parameter $\alpha > 0$ "governs the tradeoff between tree size and its goodness of fit to the data" (Hastie et al., 2001). Large values of α result in smaller tree T_α and conversely for smaller values of α. As the notation suggests, with $\alpha = 0$ the solution is the full tree T_0.

To find T_α we use weakest link pruning: we successively collapse the internal node that produces the smallest per-node increase in $R(T)$, and continue until we produce the single-node (root) tree. This gives a (finite) sequence of subtrees, and one can show this sequence must contain T_α, see Breiman et al. (1984) and Ripley (1996) for details. Estimation of α ($\hat{\alpha}$) is achieved by five- or ten-fold cross-validation. Our final tree is then denoted as $T_{\hat{\alpha}}$.

It follows that, in CART and related algorithms, classification and regression trees are produced from data in two stages. In the first stage, a large initial tree is produced by splitting one node at a time in an iterative, greedy fashion. In the second stage, a small subtree of the initial tree is selected, using the same data set. Whereas the splitting procedure proceeds in a top-down fashion, the second stage, known as pruning, proceeds from the bottom-up by successively removing nodes from the initial tree.

Theorem 1 (Brieman et al., 1984, Section 3.3) *For any value of the complexity parameter α, there is a unique smallest subtree of T_0 that minimizes the cost-complexity.*

Theorem 2 (Zhang & Singer, 1999, Section 4.2) *If $\alpha_2 > \alpha_1$, the optimal sub-tree corresponding to α_2 is a subtree of the optimal subtree corresponding to α_1.*

More general, suppose we end up with m thresholds, $0 < \alpha_1 < \alpha_2 < \ldots < \alpha_m$ and let $\alpha_0 = 0$. Also, let corresponding optimal subtrees be $\{T_{\alpha_0}, T_{\alpha 1}, T_{\alpha 2}, \ldots, T_{\alpha m}\}$, then

$$T_{\alpha_0} \succ T_{\alpha 1} \succ T_{\alpha 2} \succ \ldots \succ T_{\alpha m} \tag{9}$$

where $T_{\alpha_0} \succ T_{\alpha 1}$ means that $T_{\alpha 1}$ is a subtree of T_{α_0}. These are called nested optimal subtrees.

3. Decision Tree for Censored Survival Data

Survival analysis is the phrase used to describe the analysis of data that correspond to the time from a well-defined time origin until the occurrence of some particular events or end-points. It is important to state what the event is and when the period of observation starts and finish. In medical research, the time origin will often correspond to the recruitment of an individual into an experimental study, and the end-point is the death of the patient or the occurrence of some adverse events. Survival data are rarely normally distributed, but are skewed and comprise typically of many early events and relatively few late ones. It is these features of the data that necessitate the special method survival analysis.

The specific difficulties relating to survival analysis arise largely from the fact that only some individuals have experienced the event and, subsequently, survival times will be unknown for a subset of the study group. This phenomenon is called censoring and it may arise in the following ways: (a) a patient has not (yet) experienced the relevant outcome, such as relapse or death, by the time the study has to end; (b) a patient is lost to follow-up during the study period; (c) a patient experiences a different event that makes further follow-up impossible. Generally, censoring times may vary from individual to individual. Such censored survival time underestimated the true (but unknown) time to event.

Visualising the survival process of an individual as a time-line, the event (assuming it is to occur) is beyond the end of the follow-up period. This situation is often called right censoring. Most survival data include right censored observation.
In many biomedical and reliability studies, interest focuses on relating the time to event to a set of covariates. Cox proportional hazard model (Cox, 1972) has been established as the major framework for analysis of such survival data over the past three decades. But, often in practices, one primary goal of survival analysis is to extract meaningful subgroups of patients determined by the prognostic factors such as patient characteristics that are related to the level of disease. Although proportional hazard model and its extensions are powerful in studying the association between covariates and survival times, usually they are problematic in prognostic classification. One approach for classification is to compute a risk score based on the estimated coefficients from regression methods (Machin et al., 2006). This approach, however, may be problematic for several reasons. First, the definition of risk groups is arbitrary. Secondly, the risk score depends on the correct specification of the model. It is difficult to check whether the model is correct when many covariates are involved. Thirdly, when there are many interaction terms and the model becomes complicated, the result becomes difficult to interpret for the purpose of prognostic classification. Finally, a more serious problem is that an invalid prognostic group may be produced if no patient is included in a covariate profile. In contrast, DT methods do not suffer from these problems.
Owing to the development of fast computers, computer-intensive methods such as DT methods have become popular. Since these investigate the significance of all potential risk factors automatically and provide interpretable models, they offer distinct advantages to analysts. Recently a large amount of DT methods have been developed for the analysis of survival data, where the basic concepts for growing and pruning trees remain unchanged, but the choice of the splitting criterion has been modified to incorporate the censored survival data. The application of DT methods for survival data are described by a number of authors (Gordon & Olshen, 1985; Ciampi et al., 1986; Segal, 1988; Davis & Anderson, 1989; Therneau et al., 1990; LeBlanc & Crowley, 1992; LeBlanc & Crowley, 1993; Ahn & Loh, 1994; Bacchetti & Segal, 1995; Huang et al., 1998; Keleş & Segal, 2002; Jin et al., 2004; Cappelli & Zhang, 2007; Cho & Hong, 2008), including the text by Zhang & Singer (1999).

4. Decision Tree for Multivariate Censored Survival Data

Multivariate survival data frequently arise when we faced the complexity of studies involving multiple treatment centres, family members and measurements repeatedly made on the same individual. For example, in multi-centre clinical trials, the outcomes for groups of patients at several centres are examined. In some instances, patients in a centre might exhibit similar responses due to uniformity of surroundings and procedures within a centre. This would result in correlated outcomes at the level of the treatment centre. For the situation of studies of family members or litters, correlation in outcome is likely for genetic reasons. In this case, the outcomes would be correlated at the family or litter level. Finally, when one person or animal is measured repeatedly over time, correlation will most definitely exist in those responses. Within the context of correlated data, the observations which are correlated for a group of individuals (within a treatment centre or a family) or for

one individual (because of repeated sampling) are referred to as a cluster, so that from this point on, the responses within a cluster will be assumed to be correlated.
Analysis of multivariate survival data is complex due to the presence of dependence among survival times and unknown marginal distributions. Multivariate survival times frequently arise when individuals under observation are naturally clustered or when each individual might experience multiple events. A successful treatment of correlated failure times was made by Clayton and Cuzik (1985) who modelled the dependence structure with a frailty term. Another approach is based on a proportional hazard formulation of the marginal hazard function, which has been studied by Wei et al. (1989) and Liang et al. (1993). Noticeably, Prentice et al. (1981) and Andersen & Gill (1982) also suggested two alternative approaches to analyze multiple event times.
Extension of tree techniques to multivariate censored data is motivated by the classification issue associated with multivariate survival data. For example, clinical investigators design studies to form prognostic rules. Credit risk analysts collect account information to build up credit scoring criteria. Frequently, in such studies the outcomes of ultimate interest are correlated times to event, such as relapses, late payments, or bankruptcies. Since DT methods recursively partition the predictor space, they are an alternative to conventional regression tools.
This section is concerned with the generalization of DT models to multivariate survival data. In attempt to facilitate an extension of DT methods to multivariate survival data, more difficulties need to be circumvented.

4.1 Decision tree for multivariate survival data based on marginal model

DT methods for multivariate survival data are not many. Almost all the multivariate DT methods have been based on between-node heterogeneity, with the exception of Molinaro et al. (2004) who proposed a general within-node homogeneity approach for both univariate and multivariate data. The multivariate methods proposed by Su & Fan (2001, 2004) and Gao et al. (2004, 2006) concentrated on between-node heterogeneity and used the results of regression models. Specifically, for recurrent event data and clustered event data, Su & Fan (2004) used likelihood-ratio tests while Gao et al. (2004) used robust Wald tests from a gamma frailty model to maximize the between-node heterogeneity. Su & Fan (2001) and Fan et al. (2006) used a robust log-rank statistic while Gao et al. (2006) used a robust Wald test from the marginal failure-time model of Wei et al. (1989).
The generalization of DT for multivariate survival data is developed by using goodness of split approach. DT by goodness of split is grown by maximizing a measure of between-node difference. Therefore, only internal nodes have associated two-sample statistics. The tree structure is different from CART because, for trees grown by minimizing within-node error, each node, either terminal or internal, has an associated impurity measure. This is why the CART pruning procedure is not directly applicable to such types of trees. However, the split-complexity pruning algorithm of LeBlanc & Crowley (1993) has resulted in trees by goodness of split that has become well-developed tools.
This modified tree technique not only provides a convenient way of handling survival data, but also enlarges the applied scope of DT methods in a more general sense. This is especially for those situations where defining prediction error terms is relatively difficult, so growing trees by a two-sample statistic, together with the split-complexity pruning, offers a feasible way of performing tree analysis.

The DT procedure consists of three parts: a method to partition the data recursively into a large tree, a method to prune the large tree into a subtree sequence, and a method to determine the optimal tree size.
In the multivariate survival trees, the between-node difference is measured by a robust Wald statistic, which is derived from a marginal approach to multivariate survival data that was developed by Wei et al. (1989). We used split-complexity pruning borrowed from LeBlanc & Crowley (1993) and use test sample for determining the right tree size.

4.1.1 The splitting statistic

We consider n independent subjects, each subject have K potential types or number of failures. If there are an unequal number of failures within the subjects, then K is the maximum. We let $T_{ik} = \min(Y_{ik}, C_{ik})$ where Y_{ik} = time of the failure in the ith subject for the kth type of failure and C_{ik} = potential censoring time of the ith subject for the kth type of failure with $i = 1,\ldots,n$ and $k = 1,\ldots,K$. Then $\delta_{ik} = I(Y_{ik} \le C_{ik})$ is the indicator for failure and the vector of covariates is denoted by $\mathbf{Z}_{ik} = (Z_{1ik}, \ldots, Z_{pik})^T$.
To partition the data, we consider the hazard model for the ith unit for the kth type of failure, using the distinguishable baseline hazard as described by Wei et al. (1989), namely

$$\lambda_{ik}(t) = \lambda_{0k}(t)\exp\left(\beta \cdot I(Z_{ik} < c)\right) \tag{10}$$

where the indicator function $I(Z_{ik} < c)$ equals 1 if $Z_{ik} < c$ and 0 otherwise, which corresponds to a split, say s, based on a continuous covariate Z_j ($j = 1,..,p$). If the covariate is categorical, then $I(Z_{ik} \in A)$ for any subset A of its categories need to be considered.
Parameter β is estimated by maximizing the partial likelihood. If the observations within the same unit are independent, the partial likelihood functions for β for the distinguishable baseline model (10) would be,

$$L(\beta) = \prod_{i=1}^{n}\prod_{k=1}^{K}\left\{\frac{\exp\left(\beta \cdot I(Z_{ik} < c)\right)}{\sum_{j=1}^{n} I(T_{jk} \ge T_{ik})\exp\left(\beta \cdot I(Z_{jk} < c)\right)}\right\} \tag{11}$$

Since the observations within the same unit are not independent for multivariate failure time, we refer to the above functions as the pseudo-partial likelihood.
The estimator $\hat{\beta}$ can be obtained by maximizing the likelihood by solving $u(\beta) = \frac{\partial \log L(\beta)}{\partial \beta} = 0$. Wei et al. (1989) showed that $\sqrt{n}(\beta - \hat{\beta})$ is normally distributed with mean 0. However the usual estimate, $a^{-1}(\beta)$, for the variance of $\hat{\beta}$, where

$$a(\beta) = -\left(\frac{1}{n}\right)\frac{\partial^2 \log L(\beta)}{\partial \beta^2}\bigg|_{\beta=\hat{\beta}} \tag{12}$$

is not valid. We refer to $a^{-1}(\beta)$ as the naïve estimator. Wei et al. (1989) showed that the correct estimated (robust) variance estimator of $\sqrt{n}\left(\beta - \hat{\beta}\right)$ is

$$d(\beta) = a^{-1}(\beta)b(\beta)a^{-1}(\beta) \tag{13}$$

where $b(\beta)$ is weight and $d(\beta)$ is often referred to as the robust or sandwich variance estimator. Hence, the robust Wald statistic corresponding to the null hypothesis $H_0 : \beta = 0$ is

$$\chi_W^2 = \frac{\hat{\beta}^2}{d(\beta)} \tag{14}$$

4.1.2 Tree growing

To grow a tree, the robust Wald statistic is evaluated for every possible binary split of the predictor space Z. The split, s, could be of several forms: splits on a single covariate, splits on linear combinations of predictors, and boolean combination of splits. The simplest form of split relates to only one covariate, where the split depends on the type of covariate whether it is ordered or nominal covariate.

The "best split" is defined to be the one corresponding to the maximum robust Wald statistic. Subsequently the data are divided into two groups according to the best split.

Apply this splitting scheme recursively to the learning sample until the predictor space is partitioned into many regions. There will be no further partition to a node when any of the following occurs:

1. The node contains less than, say 10 or 20, subjects, if the overall sample size is large enough to permit this. We suggest using a larger minimum node size than used in CART where the default value is 5;
2. All the observed times in the subset are censored, which results in unavailability of the robust Wald statistic for any split;
3. All the subjects have identical covariate vectors. Or the node has only complete observations with identical survival times. In these situations, the node is considered as 'pure'.

The whole procedure results in a large tree, which could be used for the purpose of data structure exploration.

4.1.3 Tree pruning

Let T denotes either a particular tree or the set of all its nodes. Let S and $\tilde{T}$ denote the set of internal nodes and terminal nodes of T, respectively. Therefore, $T = \tilde{T} + S$. Also let $|\cdot|$ denotes the number of nodes. Let $G(h)$ represents the maximum robust Wald statistic on a particular (internal) node h. In order to measure the performance of a tree, a split-complexity measure $G_\alpha(T)$ is introduced as in LeBlanc and Crowley (1993). That is,

$$\begin{aligned} G_\alpha(T) &= G(T) - \alpha|S| \\ &= \sum_{h \in S} G(h) - \alpha|S| \end{aligned} \tag{15}$$

where the number of internal nodes, $|S|$, measures complexity; $G(T)$ measures goodness of split in T; and the complexity parameter α acts as a penalty for each additional split.
Start with the large tree T_0 obtained from the splitting procedure. For any internal node h of T_0, i.e. $h \in S_0$, a function $g(h)$ is defined as

$$g(h) = \frac{G(T_h)}{|S_h|} \tag{16}$$

where T_h denotes the branch with h as its root and S_h is the set of all internal nodes of T_h. Then the 'weakest link' $\bar{h}_0$ in T_0 is the node such that

$$g(\bar{h}_0) = \min_{h \in S_0} g(h) \tag{17}$$

Let $\alpha_1 = g(\bar{h}_0)$ and T_1 be the subtree after pruning off the branch $T_{\bar{h}_0}$. In addition, let $\alpha_2 = g(\bar{h}_1) = \min_{h \in S_1} g(h)$ and T_2 be the tree after pruning off the branch $T_{\bar{h}_1}$. Repeating this procedure leads to a nested sequence of subtrees $T_M \prec \ldots \prec T_m \prec T_{m-1} \prec \ldots \prec T_1 \prec T_0$ where T_M is the root node, and the sequence $\infty = \alpha_M > \ldots > \alpha_m > \alpha_{m-1} > \ldots > \alpha_1 > \alpha_0 = 0$.
It is important to note that α_m's are an increasing sequence. And for any α such that $\alpha_m \le \alpha \le \alpha_{m+1}$, in particular, the geometric mean of α_m, $\alpha'_m = \sqrt{\alpha_m \alpha_{m+1}}$. It follows that $T(\alpha) = T(\alpha_m) = T(\alpha'_m) = T_m$. This implies that we can get the best pruned subtree for any penalty α from the pruning algorithm.

4.1.4 The best-sized tree based on test sample

Now we need to select one or several appropriately sized trees from the nested sequence. Several methods have been suggested for this purpose, one of them is test sample method.
When the sample size is large enough, the test sample is the preferred method of determining the right tree size. To do this, the whole sample is divided into two parts: a learning sample L_1 and a test sample L_2. Usually, the proportion is 2:1.
A large tree T_0 is grown and pruned to obtain a nested sequence of subtrees using the learning sample L_1. Then the test sample L_2 is sent down along the large tree T_0 and the splitting statistics, $G(h)$, are recalculated for each internal node, $h \in S$, using the validation sample. The tree that maximizes the split-complexity measure $G_{\alpha_c}(T)$ is chosen as the best-sized subtree, where the constant penalty α_c is chosen for each split. It has been suggested by LeBlanc and Crowley (1993) that α_c be typically chosen such that $2 \le \alpha_c \le 4$, where $\alpha_c = 2$ is in the spirit of the AIC criterion and $\alpha_c = 4$ corresponds roughly to the 0.05 significance level for a split under the $\chi^2_{(1)}$ curve.

Finally a marginal Kaplan-Meier survival curve is prepared separately (i.e., for each type of failure) for all groups resulted from the best-sized tree. For example, if the best-sized trees classified patients into three groups, then we prepare three corresponding marginal Kaplan-Meier survival curves for each type of failure.

4.1.5 Application: Bladder cancer data

In this section, we shall illustrate the proposed methods using the well-known bladder tumour cancer data reported by Byar (1980). The data were from a randomized clinical trial conducted by the Veterans Administration Co-operative Urological Group between 1971 and 1976 and consisted of 117 patients with superficial bladder tumours. The tumors were removed transurethrally, and patients were then randomly assigned to one of three treatments: placebo, pyridoxine (Vitamin B6), or intravesical thiotepa (triethylenetriphosphamide). Thiotepa is a member of the class of alkylating agents, which were among the first anticancer drugs used. Alkylating agents are highly reactive and bind to certain chemical groups found in nucleic acids. These compounds inhibit proper synthesis of DNA and RNA, which leads to apoptosis or cell death. However, since alkylating agents cannot discriminate between cancerous and normal cells, both types of cells will be affected by this therapy. For example, normal cells can become cancerous due to alklyating agents. Thus, thiotepa is a highly cytotoxic compound and can potentially have adverse effects. Consequently, the effects of thiotepa on cancer recurrence and death are not obvious (Ghosh & Lin, 2000).

Treatment was aimed at preventing bladder cancer recurrence following the removal of superficial bladder tumours. Patients were examined every 3 months for recurrence of tumour and any new tumors were removed. We used the version of data presented in original paper by Wei et al. (1989) which is only available for the placebo and the thiotepa groups. There were 38 patients in the thiotepa group and 48 placebo patients. The outcome variable was number of months to the event since last tumour occurrence. Patients are censored when they die, immediately after their fourth event or when the end of the study is reached. Besides the treatment the number of initial tumours and diameter of the largest initial tumour were also recorded for each patient. Particularly, the number of initial tumours ranged from 1 to 8 with the respective counts of patients equal to 50, 11, 10, 4, 5, 2, 0, and 3.

Figure 2. shows the best-sized survival tree based on robust Wald splitting. At each level of the tree, we show the best splitter (covariate with cutpoint), and the corresponding robust Wald split statistic. A square denotes terminal nodes in the tree. Beneath each terminal node, n denotes the number of patients.

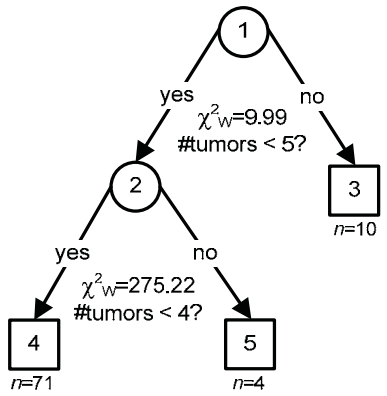

Fig. 2. Survival tree based on robust Wald splitting.

The root node was split by the number of initial tumours, with the best cutoff fewer than five versus at least five initial tumours (χ^2_W = 9.99). The subgroup with at least five initial tumours formed terminal node 3. On the opposite side of the tree, the subgroup with fewer than five positive nodes was next split by number of initial tumours again (best cutoff <4 vs. ≥4; χ^2_W = 275.22). None of these subgroups were further split and formed terminal nodes 4 and 5 in the tree. Hence, best-sized tree formed three groups of patients.

Prognostic grouping of the patients was based on the terminal nodes in Figure 2. We chose survival probability to first and second recurrence represented by marginal Kaplan-Meier curve as a measure of prognosis, and ranked each terminal node in the tree according to that measure. The survival probability of first and second recurrence for three groups of patients is presented in Figure 3. Among these three groups, patient with at least five initial tumours have the poorest prognosis, that is, they are more likely to develop recurrence. Patient with at most three initial tumours have the best prognosis, because they are less likely to develop recurrence.

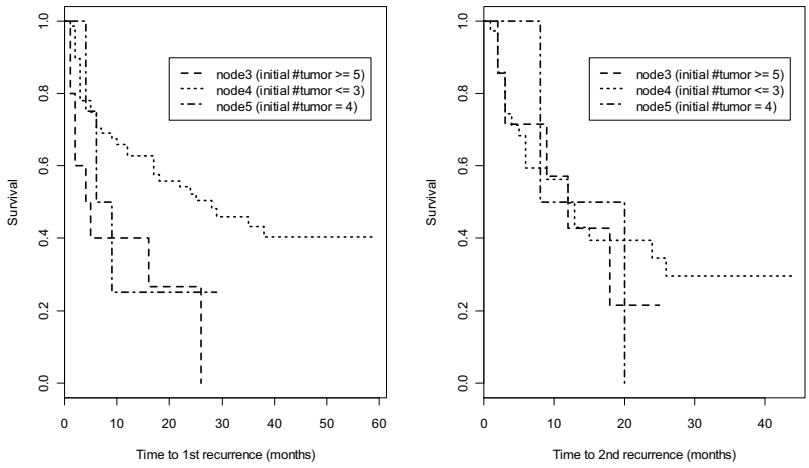

Fig. 3. Survival probability to the first (left) and second (right) recurrence for the three prognostic groups derived from terminal nodes in Figure 2.

4.2 Decision tree for multivariate survival data based on frailty model

Quite the opposite to marginal models, frailty models directly account for and estimate within subject correlation. A parameter estimate of within subject propensity for events is obtained directly.

The random effects approach to frailty models involves the assumption that there are unknown factors within a subject causing similarity (homogeneity) in failure times within the subject and thus differences (heterogeneity) in failure times between different subjects. The reason such factors are referred to as unknown is that if they were known to the investigator, they could be included in the analysis, resulting in independence within a subject. Frailty modeling (known as such because it examines the tendency for failures/recurrences within a subject at similar times, or experience similar frailties) involves specification of independence within a subject, conditional on an unobservable frailty v_i.

This frailty for the ith subject is incorporated (conditionally) into the proportional hazard function previously examined as follows

$$\lambda_{ik}(t \mid v_i) = v_i \lambda_0(t) \exp(\beta Z_{ik}) \tag{18}$$

where the number of failure for the ith subject may be different, i.e. $k = 1,\ldots,K_i$.

4.2.1 The splitting statistic and tree growing

In this method, the splitting rule is defined as Wald test statistic evaluating covariate effect based on frailty model. Thus, we create tree by maximizing between-node separation.
To split a node, a splitting function needs to be evaluated at every cutoff point. We consider the following model

$$\lambda_{ik}(t \mid v_i) = v_i \lambda_0(t) \exp(\beta \cdot I(Z_{ik} < c)) \tag{19}$$

The parameter β corresponds to the effect of separation between two daughter nodes induced by cutoff point c which will be estimated by penalized likelihood method. The penalized likelihood is developed by re-parameterizing model (19) as follows (Therneau & Grambsch, 2000):

$$\lambda_{ik}(t \mid \gamma) = \lambda_0(t) \exp(\beta \cdot I(Z_{ik} < c) + X\gamma) \tag{20}$$

where $\gamma = \log(v) = \{\gamma_1, \gamma_2, \ldots, \gamma_n\}$ is the vector of parameters for the re-parameterized frailty, and X is the corresponding design matrix, which is a N by n matrix, where $N = \sum_{i=1}^{n} K_i$, such that $X_{kl} = 1$ when kth failure belongs to lth subject and $X_{kl} = 0$ otherwise, with $k = 1,\ldots, N$ and $l = 1,2,\ldots, n$. With this parameterization, model (20) has a similar structure as the classical Cox model. Let θ be the index parameter of the frailty distribution in the node. Then, the penalized log-likelihood can be expressed as

$$PPL = PL(\beta, \gamma; \text{data}) - g(\gamma; \theta) \tag{21}$$

where $PL(\beta,\gamma;\text{data})$ is the partial likelihood and $g(\gamma;\theta)$ is a penalty function. Specifically, $PL(\beta,\gamma;\text{data})$ is the usual partial likelihood for β and γ,

$$PL(\beta, \gamma; \text{data}) = \sum_{i=1}^{N} \int_0^{\infty} \left[W_i(t)(I(z_i < c)\beta + X_i\gamma) - \log\left\{ \sum_{k \in R(t_i)} W_k(t) \exp(I(z_k < c)\beta + X_k\gamma) \right\} \right] dN_i(t) \tag{22}$$

where X_i is the ith row of design matrix which corresponds to the ith recurrence in node, $W_i(t)$ is an indicator variables such that $W_i(t)=1$ when the item is at risk in time t_i and 0 otherwise, and $R(t_i)$ is the risk set in time t_i. It has been shown (Therneau & Grambsch, 2000) that the following penalty function will give exactly the same solution as EM algorithm (Klein, 1992) when the frailty has a gamma distribution indexed by parameter θ,

$$g(\gamma;\theta) = -\frac{1}{\theta}\sum_{i=1}^{n}[\gamma_i - \exp(\gamma_i)] \tag{23}$$

The parameters β and γ can be estimated by solving the score functions and parameter θ is estimated by maximizing the profile log-likelihood (Therneau & Grambsch, 2000). In turn, following Gray (1992), the variance covariance matrix is $V(\beta;\gamma) = H^{-1}$ and

$$H = H(\beta,\gamma) = A + \begin{pmatrix} 0 & 0 \\ 0 & g'' \end{pmatrix} \tag{24}$$

where A is the second derivative matrix of usual partial likelihood with respect to β and γ, g'' is the second derivative of penalty function with respect to γ. The first diagonal element of H^{-1} corresponds to the variance of β (Therneau & Grambsch, 2000; Gray, 1992). Comparing to EM algorithm, penalized likelihood method is computationally faster and has been incorporated into standard statistical packages.

The splitting function is defined as $\chi^2 = \hat{\beta}^2 / \hat{\text{var}}(\beta)$, where $\hat{\beta}$ and $\hat{\text{var}}(\beta)$ are penalized likelihood estimator and its estimated variance. In summary, when a tree is constructed, a conditional proportional hazard structure given frailty is assumed within each node. The splitting function is evaluated at each allowable split, and the best cutoff point c^* is chosen corresponding to the maximum Wald statistic. This process is applied recursively until all the nodes cannot be further split. That is, the covariate space within the node becomes homogeneous or only a few failures are left within the node. Consequently, the growing procedure leads to a large initial tree, denoted by T_0.

4.2.2 Tree pruning and the best-sized tree selection based on test sample

The data in a node are sent to one of two child nodes by a split point or split set selected. This procedure is repeated until a certain criterion is met. A stopping rule might not detect significant splits which occur at later nodes. To avoid this possibility, a pruning technique can be applied to eliminate some insignificant nodes after splitting as many times as possible until each node has fewer than a pre-specified number of cases.

The prediction error of the maximal tree is usually larger than that of a parsimonious tree when estimated by an independent sample. Thus, some nodes need to be pruned. We adopt the split-complexity pruning technique of LeBlanc & Crowley (1993), defining Wald statistic as the goodness of split of a tree. The approach is to find a tree T maximizing the split-complexity as follows:

1. Given a complexity parameter $\alpha \geq 0$ and a tree T, define the split-complexity function $G_\alpha(T)$ as $G_\alpha(T) = G(T) - \alpha|S|$, where $G(T) = \sum_{h\in S} G(h)$ is the sum of the maximum Wald statistic over the internal nodes of T and $|S|$ is the number of internal nodes in T.
2. For each $\alpha \geq 0$, there exists a tree maximizing the split-complexity. If $\alpha = 0$, the tree maximizing the split-complexity is the maximal tree. The larger the α, the smaller the tree maximizing the split-complexity.
3. Let h be any node and T_h be the branch of a tree T with root node h, then we define a function $g(h) = G(T_h)/|S_h|$, where S_h is the set of all internal nodes of T_h.

4. Prune branches at node $\bar{h}_0$ for which $g(\bar{h}_0) = \min_{h \in S_0} g(h)$.
5. Repeat the above steps to obtain a nested sequence of trees, $T_0 \succ T_1 \succ \ldots \succ T_{m-1} \succ T_m \succ \ldots \succ T_M$, where T_0 is the maximal tree. T_m is a tree with some branches of T_{m-1} pruned, and T_M is the trivial tree with only the root node.

The final tree is selected by evaluating the sequence of maximal split-complexity trees. However, the goodness of split of the trees is over-estimated by the learning sample. That is, the estimation of the goodness of split by the learning sample is too optimistic. Thus, if an independent test sample exists, the final tree can be selected as follows:

1. Let $T_0 \succ T_1 \succ \ldots \succ T_{m-1} \succ T_m \succ \ldots \succ T_M$ be the pruned sequence of trees obtained from a learning sample L_1. Pour a test sample L_2 down each tree.
2. Estimate the split-complexity $G(T_m)$ of each tree T_m with the test sample L_2.
3. Select the tree with the largest value of $G(T_m)$.

4.2.3 Application: Chronic granulomatous disease (CGD) data

In this section, we illustrate the DT multivariate survival data based on frailty model by analyzing a CGD data in Therneau & Grambsch (2000). CGD is a heterogeneous group of uncommon inherited disorder characterized by recurrent pyogenic infections that usually begin early in life and may lead to death in childhood. Interferon gamma is a principal macrophage-activating factor shown to partially correct the metabolic defect in phagocytes, and for this reason it was hypothesized that it would reduce the frequency of serious infections in patients with CGD. In 1986, Genentech Inc. conducted a randomized, double-blind, manized interferon gamma (rIFN-g) or placebo three times daily for a year. The primary endpoint of the study was the time to the first serious infection. However, data were collected on all serious infections until cessation of follow up, which occurred near day 400 for most patients. Thirty of the group had at least one serious infection. The total number of infections was 56 and 20 in the placebo and treatment groups, respectively. A question is whether a distinct group of patient based on their recurrent infections exist.

Covariates include the enrolling hospital and randomization data, age, height, weight, sex, use of antibiotics or corticosteroids at the time of enrolment, and the pattern of inheritance. The data set had 203 observations on 128 subjects.

Next, we applied the proposed DT method to these data. Then we developed a tree by using all data, and the best sized tree with seven terminal nodes and its corresponding Kaplan-Meier survival curves are presented in Figure 4 and Figure 5, respectively. The method first splits on whether treatment is rIFN-g or placebo. Then age (< 5.5 versus ≥ 5.5 years) and height (< 132.9 versus ≥ 132.9 cm) are chosen as the best partitions respectively. After 6 partitions, a best sized tree with 7 terminal nodes was developed, where circles and squares represent internal nodes and terminal nodes respectively, and value χ^2 denoted the splitting function for each partition. Final tree with seven terminal nodes lead to seven risk groups: placebo with age < 5.5 years (node 4), rIFN-g with height < 132.9 cm (node 6), placebo with age ≥ 5.5 years and height < 158.5 cm (node 10), rIFN-g with 132.9 ≤ height < 148 cm (node 14), rIFN-g with height ≥ 148 cm (node 15), placebo with 5.5 ≤ age < 20.5 years and height ≥ 158.5 cm (node 22), and placebo with age ≥ 20.5 years and height ≥ 158.5 cm (node 23).