The International Journal of Biostatistics

Volume 4, *Issue* 1 2008 *Article* 2

Interaction Trees with Censored Survival Data

Xiaogang Su[*] Tianni Zhou[†] Xin Yan[‡]
Juanjuan Fan[**] Song Yang[††]

[*]University of Central Florida, xiaosu@mail.ucf.edu
[†]University of Southern California, tzhou@childrensoncologygroup.org
[‡]University of Missouri - Kansas City, yanxi@umkc.edu
[**]San Diego State University, jjfan@sciences.sdsu.edu
[††]National Institutes of Health, yangso@nhlbi.nih.gov

Interaction Trees with Censored Survival Data*

Xiaogang Su, Tianni Zhou, Xin Yan, Juanjuan Fan, and Song Yang

Abstract

We propose an interaction tree (IT) procedure to optimize the subgroup analysis in comparative studies that involve censored survival times. The proposed method recursively partitions the data into two subsets that show the greatest interaction with the treatment, which results in a number of objectively defined subgroups: in some of them the treatment effect is prominent while in others the treatment may have a negligible or even negative effect. The resultant tree structure can be used to explore the overall interaction between treatment and other covariates and help identify and describe possible target populations on which an experimental treatment demonstrates desired efficacy. We follow the standard CART (Breiman, et al., 1984) methodology to develop the interaction tree structure. Variable importance information is extracted via random forests of interaction trees. Both simulated experiments and an analysis of the primary billiary cirrhosis (PBC) data are provided for evaluation and illustration of the proposed procedure.

KEYWORDS: CART, censored survival times, random forests, subgroup analysis

---

*This research was supported in part by NIH grant R03 DE016924. We also would like to thank the editor, Professor Jack Kalbfleisch, and one anonymous referee, whose insightful and constructive comments have greatly improved an initial version of this manuscript.

1. Introduction

We consider comparative studies which have censored survival times as the endpoint variable. In these studies, the main goal is to assess the effect of two or more treatments on survival. Subgroup analysis is important in determining the generalizability of the results. The investigator would like to know whether and how the treatment effect varies across subgroups induced by covariates. For example, in randomized clinical trials, practitioners and regulatory agencies would like to know whether there are subgroups of trial participants who are more (or less) likely to be helped (or harmed) by the investigational treatment. Subgroup analysis helps explore the heterogeneity of the treatment effect and extract the maximum amount of information from the available data. According to a recent survey conducted by Assmann *et al.* (2000), 70% of trials published over a three-month period in four leading medical journals included subgroup analyses.

In the current practice of subgroup analysis, the subgroups, as well as the number of subgroups to be examined, are usually prespecified by the investigator. Traditional subgroup analysis is a highly subjective process leading, potentially, to dubious results. Its limitations have been widely noted (see, e.g., Assmann, *et al.* 2000, Cook, Gebski, and Keech, 2004, Popock, *et al.* 2000, and Sleight 2000). First, it is rather blinded to an analyst in identifying the true heterogeneity structure of the effect of the investigational treatment in subjectively-predefined subgroups. One may fail to identify a subgroup of prospective interest or intentionally avoid reporting subgroups (Hahn, *et al.*, 2000) where the investigational treatment is found unsuccessful or even potentially harmful. Second, significance testing (see, e.g., Song and Chi, 2007) has been the main approach in subgroup analysis. Since there is no general guideline for selecting the number of subgroups, one has to thoroughly examine numerous possibilities to assess the treatment effect. However, a large number of subgroups inevitably leads to added difficulties concerning multiplicity and a potentially severe lack of power in testing significance. Third, it is a daunting task to clearly define subgroups *a priori*, even for the field experts. For example, Parker and Naylor (2000) reviewed 67 large randomized trials of cardiovascular pharmacotherapy published between 1980 and 1997. They found that all but five focused on single-factor subgroups solely using univariate analysis techniques and a supporting rationale for the subgroup selection was lacking for most of them.

In this paper, we propose a tree-based procedure to aid in subgroup analysis. The tree method was first considered by Morgan and Sonquist (1963). By recursively bisecting the predictor space, the hierarchical tree structure

partitions the data into meaningful groups and makes available a piecewise approximation of the underlying relationship between a response and its associated predictors. The applications of tree models have been greatly advanced in various fields especially since the development of CART (classification and regression trees) by Breiman *et al.* (1984). CART has successfully addressed tree size selection and many other related practical issues. Their idea of pruning has become and remains the current standard in determining the optimal tree size.

Applying a tree procedure to guide subgroup analysis is intuitive. Subgroup analysis involves the interaction between the treatment and covariates. Trees are well known as an excellent tool for exploring interactions, the first implementation, in fact, being referred to as Automatic Interaction Detection (AID; Morgan and Sonquist, 1963). Another attractive and unique trait of tree methods is that they group data in an optimal way. By recursively bisecting the data into two subgroups that show the greatest heterogeneity in treatment effect, we are able to optimize the subgroup analysis and make it more efficient in representing the interaction structure between the treatment and covariates. The results are a set of objectively recognized subgroups, ranging from the most effective to the least effective in terms of treatment effect. The whole procedure is data-driven and automated. The grouping strategy and the number of subgroups are automatically determined by the procedure itself. The covariates used in the partition naturally define the subgroups.

Hereafter, we will use the label "interaction trees" (IT) when referring to the proposed procedure. The rest of the paper is organized as follows. In Section 2, the IT procedure is presented in detail. Section 3 contains simulation studies designed for assessing the proposed method. In Section 4, we apply the proposed tree procedure to analyze the well-known primary billiary cirrhosis (PBC) data set. Section 5 concludes the paper with a brief discussion.

2. Tree-Structured Subgroup Analysis

We consider a study designed to assess the effect of a binary treatment on censored survival times while adjusting for a number of covariates. Let F_i and C_i be the failure time and the censoring time of the i-th case, respectively. The observed data consist of $\{(T_i, \delta_i, \mathrm{trt}_i, \mathbf{x}_i) : i = 1, 2, \ldots, n\}$, where $T_i = \min(F_i, C_i)$ is the ith observed failure time; $\delta_i = 1_{\{F_i \leq C_i\}}$ is the indicator of whether the ith case is failed or censored; trt_i is the binary treatment indicator taking values 1 or 0; and $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip}) \in \mathcal{R}^p$ is a p-dimensional

covariate vector for the ith case. We assume noninformative censoring in the sense that F_i and C_i are independent given the covariate vector $\mathbf{x}_i$.

Tree-based methods have been extended to survival analysis by many authors. These extensions are usually termed as survival trees. See, e.g., LeBlanc and Crowley (1992, 1993), Keles and Segal (2002), Su and Fan (2004), and Fan et al. (2006). However, unlike these survival trees whose goal is to facilitate a tree-structured modeling of the hazard function, our goal is to identify a tree structure (denoted by $\mathcal{T}$) that accounts for the interaction structure between the treatment and covariates. That is, we are interested in the heterogeneity structure of the treatment effect on survival across terminal nodes of $\mathcal{T}$. To construct $\mathcal{T}$, we follow the convention of CART (Breiman, *et al.*, 1984), which consists of three major steps: (1) growing a large initial tree; (2) a pruning algorithm; and (3) a validation method for determining the best tree size.

The closest work to our proposed methods includes Ciampi et al. (1995) and Negassa et al. (2005), who also developed a CART-typed tree procedure for subgroup analysis. Their final model is a stratified Cox (1972) model

$$\lambda(t \,|\, \mathbf{x}_i) = \lambda_0(t; \mathcal{T}) \, \exp\left\{\beta \cdot \mathrm{trt}_i + \boldsymbol{\gamma}' \cdot \mathbf{z}_i^{(\mathcal{T})} \cdot \mathrm{trt}_i\right\}, \tag{1}$$

where $\lambda_0(t; \mathcal{T})$ is an unspecified baseline hazard function of time with stratification on the terminal nodes of $\mathcal{T}$, $\mathbf{z}_i^{(\mathcal{T})}$ is a dummy vector induced by the tree structure $\mathcal{T}$, and $\{\beta, \boldsymbol{\gamma}\}$ are unknown regression parameters of appropriate dimension. Our splitting criterion, pruning, and tree size selection procedures are all quite different from those papers. We shall discuss the differences and draw comparisons via simulation studies in the sections that follow.

2.1 *Growing a Large Initial Tree*

We first consider a single binary split, say, s, of the data. This split is induced by a threshold on a predictor X_j. If X_j is continuous, then the binary question whether $X_j \leq c$ is considered. Observations answering 'yes' go to the left child node t_L and observations answering 'no' to the right child node t_R. Here, the cutoff point c can be any constant, but in practice its choices are empirically determined by the distinct values of observed X_j.

If X_j is nominal with categories $C = \{c_1, \ldots, c_r\}$, then the form of $X_j \in A$ with $A \subset C$ induces the split. When X_j has many distinct categories, the choices of A would be massive. To reduce the computational burden, one may 'ordinal'ize X_j in a similar way as used in CART (Section 9.4; Breiman *et al.*, 1984). To do so, we estimate the treatment effect within each category of X_j and put the categories of X_j in order according to the treatment effect. Then splitting on X_j can be processed as if it was an ordinal variable.

Among all permissible splits of the data, we want to select the one such that the treatment effect differs most between its two resultant child nodes. In other words, the best split shows the greatest interaction with the treatment. A natural measure for assessing this interaction effect is a test statistic for $H_0 : \beta_3 = 0$ in the following threshold Cox (1972) proportional hazards model

$$\lambda(t \,|\, \mathbf{x}_i) = \lambda_0(t) \, \exp\left\{\beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot z_i^{(s)} + \beta_3 \cdot \mathrm{trt}_i \cdot z_i^{(s)}\right\}, \tag{2}$$

where $z_i^{(s)} = 1_{\{X_i \le c\}}$ is the indicator variable associated with split s. While several testing procedures are available for this purpose, we use the partial likelihood ratio test (PLRT) statistic,

$$G(s) = -2 \cdot (l_2 - l_1), \tag{3}$$

where l_2 is the maximized partial log-likelihood (Cox, 1975) of model (2),

$$l_2 = \sum_{i=1}^{n} \left[\delta_i \eta_i - \log\left\{\sum_{k=1}^{n} 1_{\{T_k \ge T_i\}} \exp(\eta_k)\right\}\right]$$

with $\eta_i = \beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot z_i^{(s)} + \beta_3 \cdot \mathrm{trt}_i \cdot z_i^{(s)}$, and l_1 is the maximized partial log-likelihood of the reduced model under H_0,

$$\lambda(t \,|\, \mathbf{x}_i) = \lambda_0(t) \, \exp\left\{\beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot z_i^{(s)}\right\}.$$

For a given split s, $G(s)$ follows a $\chi^2(1)$ distribution.

The best split $s^\star$ is the one that yields the maximum $G(s)$ statistic among all permissible splits; that is, $G(s^\star) = \max_s G(s)$. The data are then split according to the best split $s^\star$. The same procedure is applied to split both child nodes. Recursively doing so results in a large initial tree, denoted by $\mathcal{T}_0$.

Remark 1 Following the formulation of the Cox (1972) proportional hazards model, we have assumed the same baseline hazard function for the two child nodes induced by a single split. Alternatively, one may consider different baseline hazard functions as in Negassa et al. (2005). In this case, the splitting statistic $G(s)$ is computed as the partial likelihood ratio test statistic that compares model

$$\lambda(t \,|\, \mathbf{x}_i) = \lambda_0(t; z_i^{(s)}) \, \exp\left\{\beta_1 \cdot \mathrm{trt}_i + \beta_2 \cdot \mathrm{trt}_i \cdot z_i^{(s)}\right\}$$

with model $\lambda(t \,|\, \mathbf{x}_i) = \lambda_0(t; z_i^{(s)}) \, \exp\{\beta_1 \cdot \mathrm{trt}_i\}$. Namely, a stratification on the two child nodes is applied.

2.2 *Pruning*

The final tree could be any subtree of $\mathcal{T}_0$. However, the number of subtrees can be massive even for a moderately-sized tree. To narrow down our choices, we follow CART's pruning method to iteratively truncate the "weakest link" of the large initial tree $\mathcal{T}_0$. Since the splitting statistic G measures the heterogeneity of child nodes for each link or internal node, we may adopt the split-complexity pruning algorithm of LeBlanc and Crowley (1993). We shall briefly describe this procedure in this subsection. Interested readers are referred to CART (Breiman, *et al.*, 1984) for basic tree terminologies, such as subtree, branch, etc.

For a given tree structure $\mathcal{T}$, let $\widetilde{\mathcal{T}}$ denote the set of all terminal nodes and $|\cdot|$ denote cardinality. An interaction-complexity measure $G_\lambda(\mathcal{T})$ is introduced to evaluate the performance of an interaction tree $\mathcal{T}$:

$$G_\lambda(\mathcal{T}) = G(\mathcal{T}) - \lambda \cdot \mid \mathcal{T} - \widetilde{\mathcal{T}} \mid, \tag{4}$$

where $G(\mathcal{T}) = \sum_{h \in \mathcal{T} - \widetilde{T}} G(h)$ corresponds to the amount of heterogeneity in treatment effect represented by tree $\mathcal{T}$ and hence measures the overall goodness-of-interaction of $\mathcal{T}$, the total number of internal nodes $|\mathcal{T} - \widetilde{\mathcal{T}}|$ is used to measure the complexity of the tree, and $\lambda \geq 0$, called the complexity parameter, acts as a penalty for each added split. Given a fixed λ, a tree structure with larger $G_\lambda(\mathcal{T})$ is preferable.

The main idea of the algorithm is that, when the complexity penalty λ increases from 0, there will be a link or internal node h that first becomes ineffective, in the sense that the branch descending from h is inferior compared to h as a single terminal node. This link is then deemed as the weakest link. The threshold value for λ can be found by solving $G_\lambda(\mathcal{T}) = G(\mathcal{T}) - \lambda \cdot |\mathcal{T} - \widetilde{\mathcal{T}}| = 0$, or $\lambda = G(\mathcal{T}) / \mid \mathcal{T} - \widetilde{\mathcal{T}} \mid$.

Following such logic leads to an efficient pruning algorithm. We start with $\mathcal{T}_0$. For each internal node h of T_0, calculate the value of

$$g(h) = \frac{G(\mathcal{T}_h)}{\mid T_h - \widetilde{\mathcal{T}_h} \mid},$$

where $\mathcal{T}_h$ is the branch with h as its root and $|\mathcal{T}_h - \widetilde{\mathcal{T}_h}|$ denotes the number of internal nodes of $\mathcal{T}_h$. Then the weakest link, $h^\star$, is the node such that $g(h^\star)$ (call it λ_1) is the smallest. Let $\mathcal{T}_1$ be the subtree after pruning off the branch $\mathcal{T}_{h^\star}$ from $\mathcal{T}_0$. Apply the same procedure to prune $\mathcal{T}_1$. Repeating this procedure results in a nested sequence of subtrees $\mathcal{T}_M \prec \cdots \prec \mathcal{T}_m \prec \mathcal{T}_{m-1} \prec \cdots \prec \mathcal{T}_1 \prec \mathcal{T}_0$, where $\mathcal{T}_M$ is the tree with only the root node and $\prec$ means "is a subtree

of". Associated with this sequence of subtrees is a corresponding sequence of λ values, $\infty > \lambda_M > \cdots > \lambda_m > \lambda_{m-1} > \cdots > \lambda_1 > \lambda_0 = 0$. By Breiman *et al.* (1984) and LeBlanc and Crowley (1993), $\mathcal{T}_m$ is the smallest subtree that maximizes G_λ for any λ such that $\lambda_m \leq \lambda < \lambda_{m+1}$. In particular, this is true for the geometric mean of λ_m and λ_{m+1}, $\lambda'_m = \sqrt{\lambda_m \lambda_{m+1}}$.

2.3 *Tree Size Selection*

Now we need to select the optimally sized tree from the nested subtree sequence. Again, the split-complexity measure $G_\lambda(\mathcal{T})$ given in (4) is the yardstick for comparing these candidates. That is, tree $\mathcal{T}^\star$ is best sized if

$$G_\lambda(\mathcal{T}^*) = \max_{m=0,\ldots,M} \left\{ G(\mathcal{T}_m) - \lambda \cdot |\mathcal{T}_m - \widetilde{\mathcal{T}}_m| \right\}$$

LeBlanc and Crowley (1993) suggest that, for the purpose of size selection, λ be fixed within the range $2 \leq \lambda \leq 4$, where $\lambda = 2$ is in the spirit of the Akaike information criterion (AIC; Akaike, 1974) and $\lambda = 4$ corresponds roughly to the 0.05 significance level on the $\chi^2(1)$ curve. A similar suggestion is found in Bhansali and Downham (1977) for selecting autoregressive time series models.

However, there is one problem due to the very adaptive nature of recursive partitioning. We have already used the sample to grow and prune the tree. The goodness-of-interaction measure $G(\mathcal{T}_m)$ would be over-optimistic if computed using the same data. Thus, an "honest" estimate of $G(\mathcal{T}_m)$ is in need. This can be achieved by cross validation. In the following, the notation $G(\mathcal{L}_2; \mathcal{L}_1, \mathcal{T})$ is used to denote the validated goodness-of-split measure for tree T built using sample $\mathcal{L}_1$ and validated using sample $\mathcal{L}_2$.

When the sample size is large, a test sample method can be applied. First divide the whole data into two parts: the learning sample $\mathcal{L}_1$ and the test sample $\mathcal{L}_2$. Then grow and prune the initial tree T_0 using $\mathcal{L}_1$. At the stage of tree size determination, the goodness-of-split measure $G(\mathcal{T}_m)$ is recalculated or validated as $G(\mathcal{L}_2; \mathcal{L}_1, \mathcal{T}_m)$ using the test sample $\mathcal{L}_2$. The subtree that maximizes the validated G_λ is selected as the best-sized tree.

When the sample size is small or moderate, one has to resort to v-fold cross-validation or bootstrapping methods in order to validate $G(\mathcal{T}_m)$. We adopt a bootstrap method proposed by Efron (1983) for bias correction in the prediction problem (also see LeBlanc and Crowley, 1993). In this method, one first grows and prunes a large initial tree $\mathcal{T}_0$ using the whole data. Next, bootstrap samples $\mathcal{L}_b, b = 1, \cdots, B$, are drawn from the entire sample $\mathcal{L}$. A rough guideline for the number of bootstrap samples is $25 \leq B \leq 100$, following LeBlanc and Crowley (1993). Based on each bootstrap sample $\mathcal{L}_b$, a

tree T_0^b is grown and pruned. Let $\mathcal{T}_b(\lambda'_m), m = 1, \cdots, M$, denote the optimally pruned subtrees corresponding to the λ'_m values, where $\lambda'_m, m = 1, \cdots, M$, are geometric means of the λ_m values obtained from pruning T_0. Then the bootstrap estimator of $G(\mathcal{T}_m)$ is given by

$$G^{(B)}(\mathcal{T}_m) = G(\mathcal{L}; \mathcal{L}, \mathcal{T}(\alpha'_m)) - \frac{1}{B}\sum_{b=1}^{B}\{G(\mathcal{L}_b; \mathcal{L}_b, \mathcal{T}_b(\lambda'_m)) - G(\mathcal{L}; \mathcal{L}_b, \mathcal{T}_b(\lambda'_m))\}\,. \tag{5}$$

The rational is as follows. The first part of (5), $G(\mathcal{L}; \mathcal{L}, \mathcal{T}(\lambda'_m))$, is over-optimistic as the same sample $\mathcal{L}$ is used for both tree construction and calculation of G. The second part of (5) aims to correct this bias, being an estimate of the difference between G's when the same ($\mathcal{L}_b$ and $\mathcal{L}_b$) versus different ($\mathcal{L}_b$ and $\mathcal{L}$) samples are used for growing/pruning the tree and for recalculating G, and averaged over B bootstrap samples.

Remark 2 Ciampi et al. (1995) adopts the cost-complexity pruning algorithm of CART (Breiman et al., 1974). In their procedure, the cost is defined as the partial likelihood associated with model (1). To select the best tree size, they considered four different selection methods: cross-validation, 1SE rule, minimum AIC rule, and an elbow rule, all requiring a fit of model (1). According to the suggestion of Negassa et al. (2005), a two-step procedure should be applied to determine the best tree size: first rely on cross-validation to decide whether an interaction tree structure is necessary, and then apply the elbow rule to select the best tree size. In the elbow rule, one determines the tree size by browsing the plot of validated costs and looking for "kinks" which might be local minima, not necessarily the global minimum.

Nevertheless, the partial likelihood associated with model (1) does not exactly measure the strength of interaction accounted for by a tree structure. A higher partial likelihood score may be caused by the stratification on subgroups, instead of the interaction between treatment and subgroups. Furthermore, fitting model (1) itself could be problematic, especially when $\mathcal{T}$ has many terminal nodes. As indicated by Kalbfleisch and Prentice (Sections 4.4 and 4.7, 2002), one would encounter efficiency loss in estimation when stratification is introduced unnecessarily. This, in particular, could be a problem in the tree method of Ciampi et al. (1995) and Negassa et al. (2005) as their whole procedure relies heavily on the maximized partial likelihood associated with model (1).

Our proposal circumvents this difficulty by utilizing the maximum splitting statistic at each partition. However, a tree model similar to Negassa et al. (2005) can be readily fit by using our procedure. The only modification

necessary is to employ the alternative splitting statistic discussed in Remark 1. Note that we do not have to fit the global model (1) in our tree method. This helps avoid the potential loss of efficiency due to unnecessary or excessive stratification.

2.4 *Summarizing the Terminal Nodes*

The subgroups are determined by the terminal nodes of the best interaction tree structure. The total number of subgroups, which may be further reduced, correspond to the automatically selected best tree size. Unlike conventional subgroup analysis, these subgroups obtained from the IT procedure are mutually exclusive.

Subgroup analysis has been a highly controversial subject (see, e.g., Sleight 2000). This is mainly due to the subjective process used to determine the subgroups and the multiplicity issue that emerges from testing across many subgroups. Nevertheless, it is generally agreed that subgroup analysis should be regarded as supportive and exploratory. And it is often useful for generating new research hypotheses for future studies. As recommended by Lagakos (2006), it is best not to present p-values for within-subgroup comparisons, but rather to give an estimate of the magnitude of the treatment difference and corresponding confidence intervals. To summarize the terminal nodes in our setting, one can compute the median survival time for both treatment groups, give an estimate of the hazard ratio between treatments, and present the comparative Kaplan-Meyer survival curves within each terminal node.

Very often the treatment is expected to show homogeneous effects in some of the terminal nodes, especially those from different branches. A merging scheme would be useful in this case. In this algorithm, one computes the G statistic in equation (3) between every pair of terminal nodes. The pair showing the least heterogeneity in treatment effect are then merged together. The same procedure is executed iteratively until all the remaining subgroups show heterogeneity above a reasonable threshold. This merging scheme further reduces the number of subgroups and results in a better representation of the heterogeneity structure for the treatment effect. In the end, one can sort the final subgroups based on the effect of the investigational treatment, from the most effective to the least effective. We will illustrate this merging idea with the example presented in Section 4. We also note that Ciampi *et al.* (1995) performs merging by fitting model (1) and examining the coefficient estimates.

2.5 *Variable Importance via Random Forests*

Variable importance ranking (see, e.g., van der Laan, 2006) is another attractive feature offered by recursive partitioning. It provides excellent augmentation to tree analyses based on one single tree structure. In the context of subgroup analysis, it will help answer questions such as which covariates are important effect-modifiers for the treatment. This issue cannot be fully addressed by simply examining the splitting variables shown in one single final IT structure, as an important variable can be completely masked by other correlated ones. While there are many methods available for extracting variable importance information, we develop an algorithm analogous to the procedure used in random forests (Breiman, 2001), which is among the newest and most promising developments in this regards.

Once again we make use of the overall goodness-of-interaction measure $G(\mathcal{T})$ for tree $\mathcal{T}$. Let V_j denote the importance of the j-th covariate X_j for $j = 1, \ldots, p$. We construct random forests of interaction trees by taking B bootstrap samples $\mathcal{L}_b$, $b = 1, \ldots, B$. Different from constructing an ordinary IT structure, ITs in random forests are grown with a greedy search over only a subset of randomly selected m covariates at each split and without pruning and tree size selection. For each tree T_b, the b-th out-of-bag sample (denoted as $\mathcal{L} - \mathcal{L}_b$), which contains all observations that are not in $\mathcal{L}_b$, is sent down to compute $G(\mathcal{T}_b)$. Next, the values of the j-th covariate in the out-of-bag sample $\mathcal{L} - \mathcal{L}_b$ are randomly permuted. The permuted out-of-bag sample is run down $\mathcal{T}_b$ again to recompute $G_j(\mathcal{T}_b)$. Then the relative difference between $G(\mathcal{T}_b)$ and $G_j(\mathcal{T}_b)$ is recorded. This is done for every covariate. The procedure is repeated for B bootstrap samples. Finally, the importance V_j is the average of those relative differences over all B bootstrap samples.

The whole procedure is summarized in Algorithm 1. Because of the mechanism of constructing random forests, a covariate that is unimportant but has many levels may show up frequently in the tree structure. The out-of-bag samples are used to achieve internal validation in bootstrapping. Permutation of covariate values provides further help in reducing the potential bias in variable importance determination.

3. Simulated Studies

This section contains simulation experiments designed to evaluate the performance of the IT procedure in detecting the true interaction structure. We

generated data from the following four models:

$$\begin{aligned}
\mathbf{A} \quad \lambda(t) &= \exp\{-2 + \log(2) \cdot \text{trt} - 1.5 \cdot Z_1 + 1.5 \cdot Z_2\} \\
\mathbf{B} \quad \lambda(t) &= \exp\{-2 + \log(2) \cdot \text{trt} - 1.5 \cdot Z_1 \cdot \text{trt} + 1.5 \cdot Z_2 \cdot \text{trt}\} \\
\mathbf{C} \quad \lambda(t) &= \exp\{-2 + \log(2) \cdot \text{trt} - 3 \cdot X_1 \cdot \text{trt} + 3 \cdot X_3 \cdot \text{trt}\}, \\
\mathbf{D} \quad \lambda(t) &= \exp\{-2 + \log(2) \cdot \text{trt} - \log(2) \cdot Z_1 \cdot \text{trt} + \log(2) \cdot Z_2 \cdot \text{trt}\}
\end{aligned}$$

where $Z_1 = 1_{\{X_1 \leq 0.5\}}$ and $Z_2 = 1_{\{X_2 \in (a,c)\}}$. Each data set involves a binary treatment and four covariates X_1 to X_4. Covariates X_1 and X_3 are simulated from a discrete uniform distribution over values (0.1, 0.2, ..., 1.0) while X_2 and X_4 are nominal, each having four levels $\{a, b, c, d\}$. However, only a subset of these covariates interact with the treatment. More specifically, model A is an additive model with no interaction. A null interaction tree structure with the root node only is expected. This model helps assess the type I error or false-positive rate when using the IT procedure. models B and D involve two additive terms of thresholds on X_1 and X_2 that interact with the treatment effect but model B has a stronger interaction signal than model D. If the IT procedure works well, it is expected to select a tree structure with four terminal nodes. In model C, the original values of X_1 and X_3 interact with treatment directly. In this case, a large tree is needed to represent the interaction structure.

We assess both the test sample and bootstrap methods in selecting the optimal tree size. For the test sample method, a sample size $n = 450$ is used, 300 observations forming the learning sample $\mathcal{L}_1$ and 150 observations forming the validation sample $\mathcal{L}_2$. For the bootstrap method, a sample size $n = 300$ is used. We consider two censoring rates, 0% and 50%. Each model is examined for 100 simulation runs. And for each simulated data set, three choices of λ, $\{2, 3, 4\}$, are applied to determine the best tree size.

The relative frequencies of the final tree sizes selected by the IT procedure are reported in Table 1. The expected final tree size for each model has been highlighted in boldface. In addition to the correct tree size selection, there is also a variable selection issue involved. For example, both X_1 and X_2, but neither X_3 nor X_4, are actually involved in the IT structure for models B and D. If this is the case in a particular run, we say a 'hit' is made. Similarly, we expect to see both X_1 and X_3 and only these two variables show up in the final tree for model C and none of them for the null model A. To address this variable selection issue, we counted the frequency of 'hits'. The results are also presented in the last column of Table 1.

Table 1

Simulation Results for Assessing the Tree Procedure: frequencies of the final tree sizes identified by the interaction tree (IT) procedure in 100 runs.

Model	Sample Size	Validation Method	Censoring Rate	λ	Final Tree Size 1	2	3	4	5	6	≥ 7	Hits
A	450	test sample	0%	2	**79**	6	5	5	2	3	0	79
				3	**93**	5	2	0	0	0	0	93
				4	**96**	2	2	0	0	0	0	96
			50%	2	**77**	6	7	4	2	2	2	77
				3	**94**	3	2	1	0	0	0	94
				4	**96**	3	0	1	0	0	0	96
	300	bootstrap	0%	2	**71**	0	1	5	14	8	1	71
				3	**87**	2	3	3	3	2	0	87
				4	**94**	3	2	1	0	0	0	94
			50%	2	**72**	1	1	4	16	6	0	72
				3	**94**	1	2	2	1	0	0	94
				4	**98**	0	1	1	0	0	0	98
B	450	test sample	0%	2	0	0	4	**57**	12	11	16	68
				3	0	0	8	**70**	11	4	7	82
				4	0	0	10	**75**	9	2	4	88
			50%	2	1	6	11	**33**	14	12	23	46
				3	4	14	19	**39**	9	6	9	58
				4	9	20	29	**32**	7	2	1	62
	300	bootstrap	0%	2	0	0	1	**26**	35	30	8	42
				3	0	2	3	**52**	29	10	4	64
				4	0	2	6	**66**	20	5	1	79
			50%	2	1	0	5	**15**	39	33	7	34
				3	2	5	15	**33**	31	13	1	54
				4	3	12	24	**34**	20	7	0	59
C	450	test sample	0%	2	2	1	15	34	16	18	**14**	76
				3	2	6	23	39	13	12	**5**	83
				4	5	13	25	39	8	8	**2**	75
			50%	2	5	15	19	25	13	16	**7**	54
				3	16	17	26	23	9	6	**3**	56
				4	19	23	22	19	11	4	**2**	52
	300	bootstrap	0%	2	0	0	0	0	0	2	**98**	9
				3	0	0	0	11	5	5	**79**	33
				4	0	0	3	17	15	11	**54**	58
			50%	2	1	1	3	2	0	2	**91**	8
				3	1	3	14	10	5	4	**63**	34
				4	5	15	35	22	9	9	**5**	54
D	450	test sample	0%	2	23	29	16	**13**	6	11	2	17
				3	42	33	12	**9**	1	2	1	13
				4	55	31	12	**1**	0	1	0	10
			50%	2	49	19	10	**7**	9	5	1	7
				3	69	17	6	**4**	4	0	0	5
				4	90	8	0	**2**	0	0	0	2
	300	bootstrap	0%	2	9	14	9	**15**	19	21	13	19
				3	30	11	6	**11**	16	21	5	12
				4	52	15	11	**9**	6	4	3	13
			50%	2	33	15	12	**11**	10	11	8	15
				3	59	5	4	**8**	12	7	5	10
				4	81	4	3	**5**	5	1	1	9

It can be seen that the IT procedure does reasonably well in detecting the true interaction structure and selecting the desired splitting variables in all models but model D. We first look at the results for the null model, model A. The results correspond roughly to the size of a statistical test or the probability of making type I errors. This is particularly important in subgroup analysis as we really do not want to identify subgroups across which the treatment effects are, in fact, the same. One of major criticisms towards conventional subgroup analysis is that one can always find something if one looks hard enough. The problem has led to Sleight's (2000) comments that subgroup analyses are 'fun to look at, but don't believe them.' The IT procedure correctly selects the null tree structure at least 71 times out of 100 runs. When $\lambda = 4$ is used, the percentage of correct selections is over 94%, which yields an empirical size of 100-94% = 6%, staying well within the acceptable level. This implies that the chance the IT procedure extracts an unsolicited interaction structure or makes false positive errors is rather small. For models B and C, the IT procedure also successfully signals the existence of interactions (*i.e.*, by selecting a non-null tree) for a majority of the runs and identifies the true final IT structure. The test sample and bootstrap methods show similar performance when working with models A and B. With model C, the bootstrap method seems to overfit a little more than the test sample method. In general, it is not surprising to see that weaker signals, heavier censoring, and smaller sample sizes lead to deteriorated performance for both methods. For example, the tree procedure clearly performs poorly for model D, which, however, can be explained by the very weak signals. Actually, with data generated from model D and of sample sizes $n = 150$ or 300, one can verify that one or both of the interaction terms are insignificant for most of the runs even if one fits the Cox proportional hazards model with terms (trt, Z_1, Z_2, $Z_1 \cdot$ trt, and $Z_2 \cdot$ trt). In other words, the interaction signal in model D is too weak to be detectable using the current simulation setting.

When comparing different complexity parameters, the results are mixed. However, $\lambda = 4$ seems to provide favorable selection for most cases considered in our simulation study. First it wins out in model A. This is important as we definitely do not want to extract interactions that actually do not exist. It also provides the best selection in model B. In model C, when the sample size is 450 and the test sample method is used, $\lambda = 3$ seems to work best in terms of a high frequency of hits. But the selection by $\lambda = 4$ provides only slightly worse results in this case. When the bootstrap method is used, the choice of $\lambda = 4$ performs best in terms of excluding spurious splits in the final tree structure, even though a greater number of trees with at least seven terminal

nodes were selected when $\lambda = 2$ or 3. This can be seen from the column of "hit" frequencies. When there is no censoring, $\lambda = 4$ selects a final tree structure that is split by X_1 and X_3 and only by them for 58% of the runs, while this percentage is only 9% for $\lambda = 2$. Based on the above simulation studies, we recommend using $\lambda = 4$ for final tree size determination.

In terms of comparison with the tree method of Ciampi et al. (1995), we first notice that their best selection method (the 1SE rule) could correctly identify the null tree structure for about 85% of the runs in both censored and noncensored cases, as reported in Negassa et al. (Page, 237; 2005). This corresponds to an empirical 'size' of 15%. We suspect this is because their procedure based on the stratified partial likelihood tends to pick up unnecessary structures that are due to stratification rather than interaction. To facilitate another comparison, we next try out the same model used in their simulations, which can be expressed as

$$\lambda(t) = \exp\left\{\log(0.33) \cdot (1 - X_1) \cdot 1_{\{X_2<40\}} \cdot \mathrm{trt} + \log(3) \cdot X_1 \cdot 1_{\{X_3>2\}} \cdot \mathrm{trt}\right\}.$$

Each data set contains $n = 600$ observations and six covariates, although only X_1, X_2, and X_3 actually modify the treatment effect. Here, X_1 and X_4 are binary 0-1 variables; $X_2, X_5 \sim \mathrm{Unif}(0, 100)$; and $X_3, X_6 \sim$ Discrete Uniform $\{1, 2, 3, 4, 5\}$. To make the comparison fair, the bootstrap method is used to construct interaction trees. We report only the results when $\lambda = 4$. Figure 1 plots the relative frequency of final tree sizes obtained from 150 simulation runs. It can be seen that our method performs better when no censoring is involved. Their best method, the elbow rule, yields only about 55% of accurate selections while ours is 67%. When the censoring rate is 50%, our results are similar to theirs, both around a 40% chance of selecting a final tree of size four. Nevertheless, the elbow rule is a somewhat subjective method.

As suggested by a referee, we made some further efforts to gain more insight into the potential optimism or bias involved in the tree procedure. Consider models B and C with no censoring. For each simulation run, we generated three independent data sets: the training or learning sample $\mathcal{L}_1$, the test sample $\mathcal{L}_2$, and the validation sample $\mathcal{L}_3$, which contain 300, 150, and 450 observations, respectively. For the final tree structure identified by the test sample method, we computed two test statistics separately using the pooled data $\mathcal{L}_1 \cup \mathcal{L}_2$ and using the validation sample $\mathcal{L}_3$: (1) the likelihood ratio test (LRT) for overall interaction; and (2) the logrank test for treatment effect within the terminal node that showed maximal treatment efficacy. Both tests are usually referred to χ^2 distributions with respective degrees of freedom (df) $|\widetilde{T}| - 1$ and 1. We recorded the resultant p-values, since the df associated with the LRT for overall

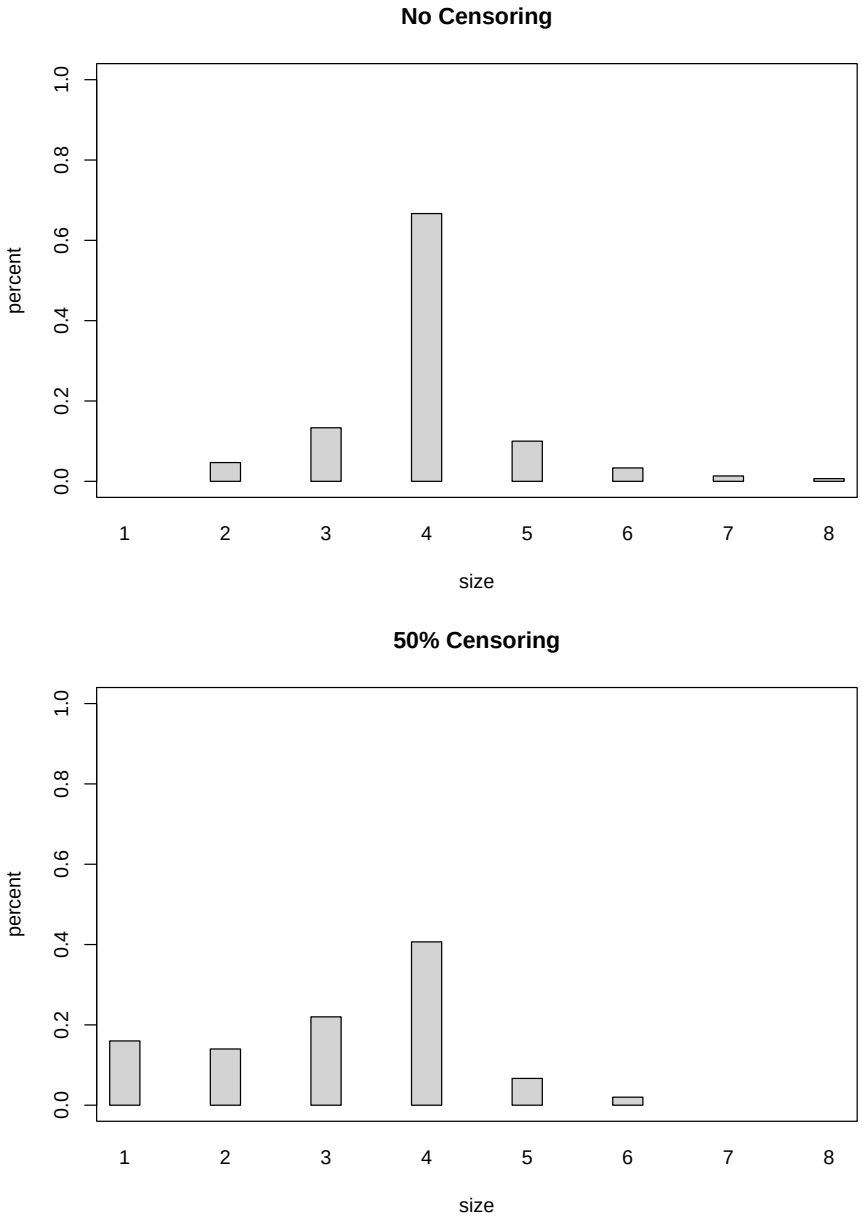

Figure 1. Comparison with Negassa et al. (p. 235; 2005): Relative Frequencies (in Percentage) of Final Tree Sizes Selected with $\lambda = 4$. Results are based on 150 runs and the sample size for each run is 600. The bootstrap method is used to determine the tree size.

Table 2

Bias Evaluation for Interaction Trees. The sample sizes for the training sample $\mathcal{L}_1$, the test sample $\mathcal{L}_2$, and the validation sample $\mathcal{L}_3$ are 300, 150, and 450, respectively. Using the pooled data $\mathcal{L}_1 \cup \mathcal{L}_2$ and the validation sample $\mathcal{L}_3$ separately, we computed the LRT for overall interaction and the logrank test for treatment effect within the subgroup showing maximal treatment efficacy and their corresponding p-values. The mean and sd of the resultant logworths, which are defined as minus logarithms of p-values with base 10, are reported out of 100 runs.

		LRT for Interaction				Logrank Max Treatment Effect			
		$\mathcal{L}_1 \cup \mathcal{L}_2$		$\mathcal{L}_3$		$\mathcal{L}_1 \cup \mathcal{L}_2$		$\mathcal{L}_3$	
Model	λ	mean	sd	mean	sd	mean	sd	mean	sd
B	2	20.666	4.206	18.907	4.513	16.788	3.687	15.947	4.464
	3	20.515	4.525	18.716	4.827	17.245	3.428	16.376	4.337
	4	20.083	4.638	18.254	4.920	17.285	3.376	16.322	4.224
C	2	20.417	4.621	16.324	4.586	14.239	4.012	12.231	4.316
	3	19.777	4.823	15.755	4.855	14.562	3.941	12.254	4.242
	4	18.670	5.658	14.889	5.609	14.449	4.381	11.948	4.611

interaction depends on the final tree size. For presentation convenience, the logworth of the p-value, which is defined as $-\log_{10}$ (p-value), was used. Note that the higher the logworth, the more significant result. Table 2 presents the mean and standard deviation of the logworths out of 100 runs. A clear inflation of logworth can be seen when the results are not based on an independent validation sample $\mathcal{L}_3$. This is particularly the case for model C when the final tree is of larger size. Also, with a smaller complexity parameter λ that leads to a larger final tree, we expected to see more inflation; nevertheless, the empirical results are mixed. Once again, we would like to remind the user to be keenly aware of the possible optimism involved in this highly adaptive procedure. When another independent sample is available, summarization of the final subgroups is best conducted based on this independent sample. When such an independent data set is not available, one should be cautious when interpreting the results.

4. An Example - The PBC Data

As an illustration, we consider data from a randomized placebo-controlled trial of the drug D-penicillamine (DPCA) for the treatment of primary billiary cirrhosis (PBC) conducted at the Mayo Clinic between 1974 and 1984 (Fleming and Harrington, 1991). Among the 312 subjects randomized to the study,

125 died by the end of follow-up. For each patient, 16 clinical, biochemical, serologic, and histologic measurements were collected. Since this is a well-known data set which has been widely studied in the statistics literature, one is referred to Fleming and Harrington (1991) and Dickson *et al.* (1985) for detailed description of the clinical background, design, variables, and related analyses.

The logrank test for assessing the effectiveness of DPCA compared to placebo yields a p-value of 0.7498. After adjusting for covariates, the p-value is 0.4654, see pp.153-162 of Fleming and Harrington (1991). Hence the study established that DPCA is not effective for the treatment of PBC.

Now we apply the proposed IT procedure to explore subgroups that account for possible heterogeneity in the effect of DPCA. We first used sample medians to impute the missing values of four variables, `triglycerides`, `serum cholesterol`, `platelets`, and `urine copper`, as they have highly skewed distributions. A similar strategy was used in Fleming and Harrington (1991).

Due to its relatively small sample size and heavy censoring (rate 59.9%), the bootstrap method was utilized. To proceed, a large initial tree was grown and pruned using the whole data set. With some restrictions on the minimum node size and the maximum tree depth, we grew an initial tree $\mathcal{T}_0$ of 10 terminal nodes. After pruning, a sequence of 5 subtrees were obtained. Thirty ($B = 30$) bootstrap samples were then generated to validate the $G(\mathcal{T}_m)$ statistic for each subtree $\mathcal{T}_m$. Figure 2 plots the validated $G_\lambda(\mathcal{T}_m)$ values versus tree sizes, as well as several intermediate measures during the process. It can be seen from Figure 2(d) that the choices of $\lambda = 3$ and $\lambda = 4$ selected the same best interaction tree at size 5 while $\lambda = 2$ selected $\mathcal{T}_0$ at size 10.

The best IT structure $\mathcal{T}^\star$ of size 5 is plotted in Figure 3, together with some related summary statistics. Next, the amalgamation algorithm was run to merge the terminal nodes of $\mathcal{T}^\star$, which resulted in three final subgroups. Table 3 summarizes the three final subgroups. The numbers of observations and deaths and the median survival time within each subgroup are included. The hazard ratio and the two-sample logrank test statistic for comparing DPCA versus the placebo are also provided. Note that one should be very cautious in interpreting their associated p-values and confidence intervals due to the very adaptive nature of the IT method. These three subgroups are then ranked as I-III according to the effectiveness of DPCA versus the placebo. These rankings are also marked next to each terminal node in Figure 3. Figure 4 plots the comparative Kaplan-Meyer survival curves for each subgroup.

The findings are interesting. First of all, the overall effect of DPCA has been deemed as insignificant. This is true for the majority of the sample, i.e.,

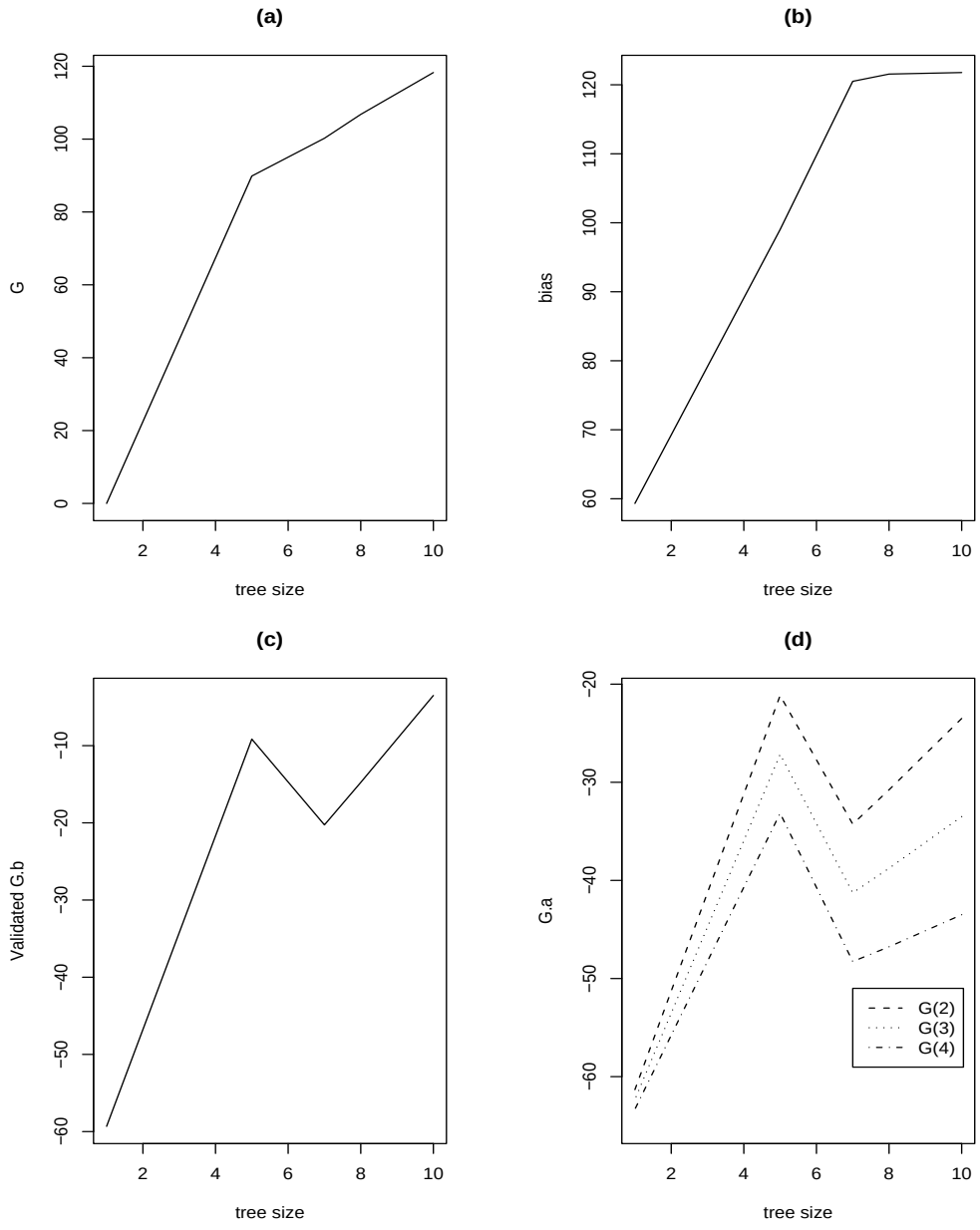

Figure 2. Final Tree Size Determination for the PBC Data via Bootstrap Resampling: (a) plot of $G(\mathcal{T}_m)$ vs. tree size; (b) plot of the estimated bias vs. tree size; (c) plot of bias-corrected $G^{(B)}(\mathcal{T}_m)$ vs. tree size; (d) plot of $G_\lambda(\mathcal{T}_m)$ vs. tree size for three different choices of λ (2, 3, and 4).

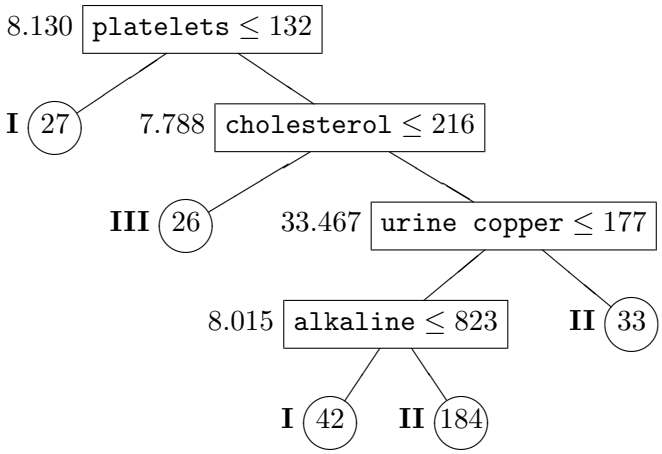

Figure 3. The Best-Sized Interaction Tree for the PBC Data. For each internal node denoted by a box, the splitting rule is given inside the box. Observations satisfying the condition proceed to the left node while observations not satisfying the condition proceed to the right node. To the left of each internal node is the PLRT statistic G. Terminal nodes are denoted by circles and ranked with Roman numerals on the left. The node size (i.e., number of observations) is given inside the circle.

Group II. However, we can see that DPCA seems to dramatically help improve survival in Subgroup I, which makes up a substantial portion, 22%, of the subjects in the trial. Patients in Subgroup I are characterized by {`platelets` $\leq$ 132} or {`platelets` $>$ 132 & `serum cholesterol` $>$ 216 & `urine copper` $\leq$ 177 & `alkaline` $\leq$ 823}. PBC is a fetal chronic liver disease of unknown cause. Until recently, effective treatments for PBC did not exist, and the approach to patients with the disease was limited to supportive care (Fleming and Harrington, 1991). This finding suggests that DPCA could be potentially useful for patients falling into the Group I category. The IT procedure also identified a subgroup, Group III, in which DPCA performs worse than the placebo. It is characterized by {`platelets` $>$ 132&`serum cholesterol` $\leq$ 216}. Group III contains only 26 individuals. Although the sample size is too small to draw any reliable conclusion, this piece of information could be useful for modifying the exclusion criteria for future study designs or in the considerations of a drug label.

To gain further insight into these subgroups, we fit a separate Cox (1972) model within each subgroup by incorporating the baseline covariates used in Fleming and Harrington (p.162, 1991). The results are included in Table 3(b).

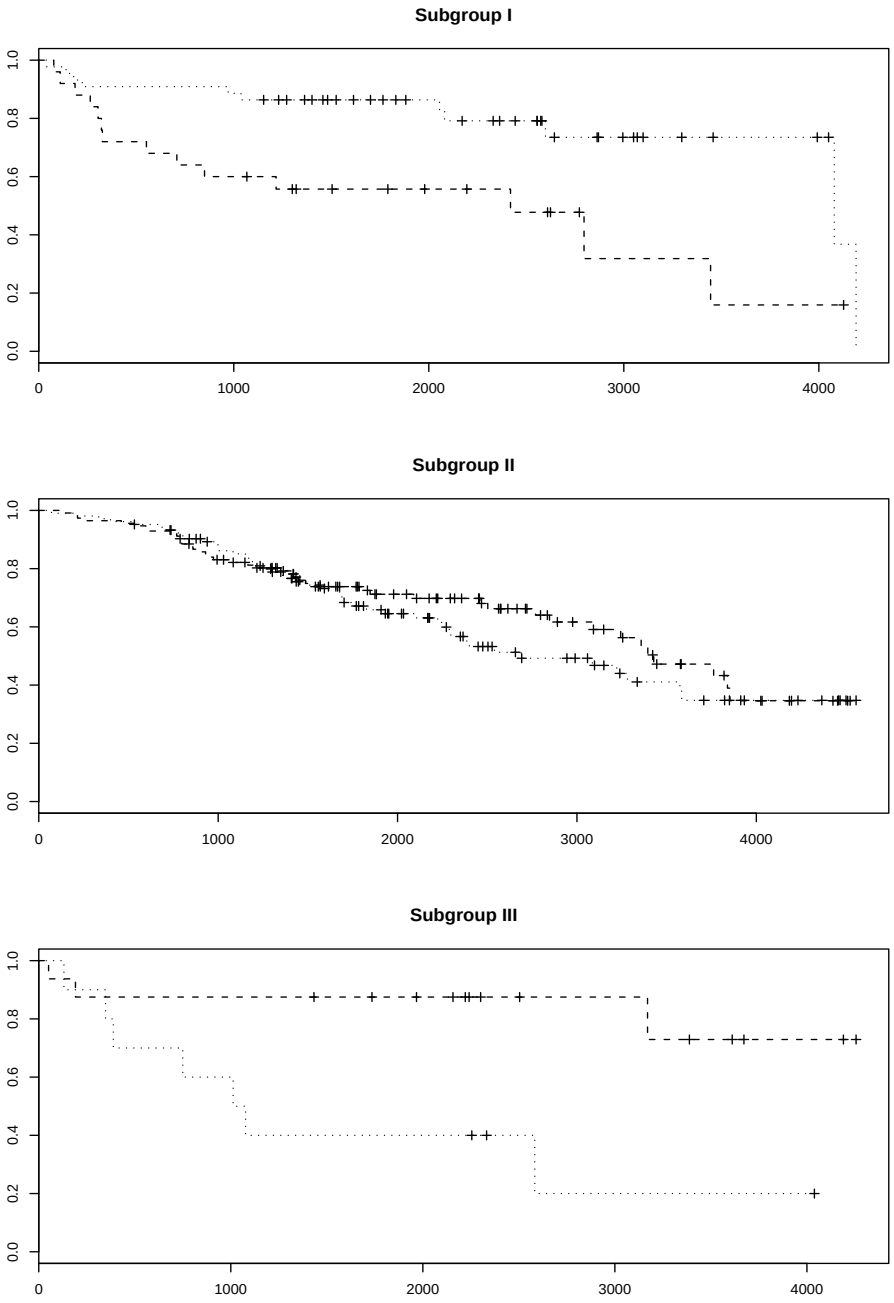

Figure 4. The Kaplan-Merier Survival Curves within Each of the Three Identified Subgroups for the PBC Data. The dotted line corresponds to the D-penicillamine (DPCA) group and the dashed line corresponds to the placebo group.

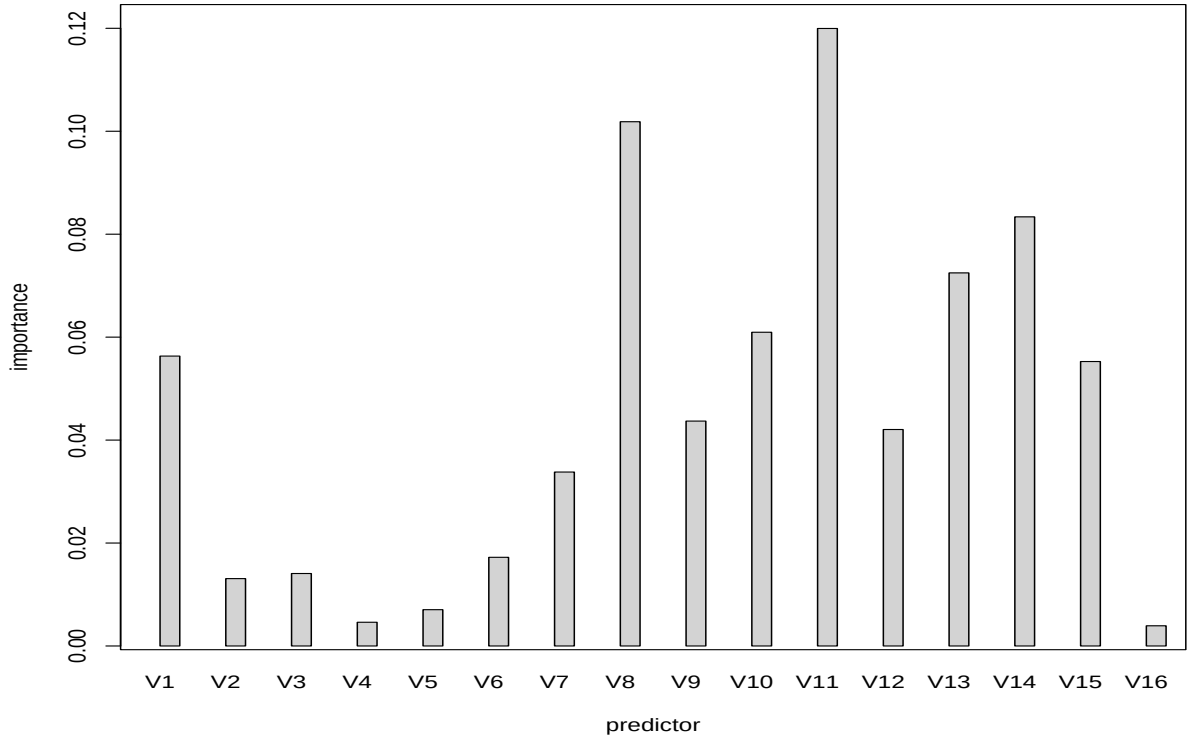

Figure 5. Variable Importance for the PBC data via Random Forests of Interaction Trees. The sixteen covariates involved here are `age` (V_1), `sex` (V_2), `ascite` (V_3), `hepatomegaly` (V_4), `spiders` (V_5), `edema` (V_6), `bilirubin` (V_7), `cholesterol` (V_8), `albumin` (V_9), `urine copper` (V_{10}), `alkaline` (V_{11}), `SGOT` (V_{12}), `triglicerides` (V_{13}), `platelets` (V_{14}), `prothrombin` (V_{15}), and `stage` (V_{16}).

Note that the treatment effect in Group III is no longer prominent after adjusting for covariates. In fact, none of the covariates was found significant in the multiple Cox (1972) model. This may again be due to the small sample size in this group.

We computed variable importance for all sixteen predictors, as plotted in Figure 5. The computation was based on 500 bootstrap samples. The results show that `alkaline` appears to be the most important effect-modifier, followed by `cholesterol`, `platelets`, `triglicerides`, and `urine copper`. This matches well with the final tree structure in Figure 3, except that `triglicerides` might have been masked out.

As a word of caution, one should always be keenly aware of not only the exploratory nature of subgroup analysis itself but also the adaptive nature of recursive partitioning. The exploratory nature of subgroup analysis entails that no decisive conclusion be drawn from the results. In terms of the adaptive nature of the IT procedure, both the greedy search scheme and the amalgamation algorithm tend to yield optimistic results. Thus, caution should be exercised in interpreting the findings in Table 3. In the case of large samples, one convenient way of undermining this optimism is to compute the figures presented in Table 3 using another independent data set. When the sample size is relatively small, which is the case for this PBC example, how to produce a more 'honest' estimate of the treatment effect within the final subgroups seems to pose additional challenges for future research.

5. Discussion

We propose a tree-based method, the IT procedure, to conduct subgroup analysis with censored survival data. Although we employ a PLRT statistic on interaction as the splitting measure, the main tree procedure does not involve any significance testing. The IT procedure provides a data-driven, objective, and automatic way to assess and explore the heterogeneity structure of the treatment effect across subgroups.

Interaction between treatment and covariates is the essence of subgroup analysis. Detecting interaction is always a challenging problem. In traditional analyses, interactions are modeled with cross-product terms, which is not efficient as interactions may occur in complicated forms. Recursive partitioning offers a nonparametric way to explore interactions. However, it handles interactions *implicitly*. Often it is still very hard to tell from a given tree structure whether interactions really exist and how variables interact with each other. The proposed IT procedure instead focuses *explicitly* on the interaction

Table 3

Summary Statistics for the Three Final Subgroups of the PBC data. Note that the median survival time in Table (a) for the placebo group in Subgroup III is not available as there are only 3 deaths out of 16 patients. In Table (b), the covariates considered are `age`*, log(*`albumin`*),* `log(bilirubin)`*,* `edema`*, and log(*`prothrombin time`*). These are the same covariates used in the best natural history model built by Fleming and Harrington (p.162, 1991).*

(**a**)

	sample	DPCA			placebo		
subgroup	size	deaths	censorings	median	deaths	censorings	median
I	69	11	33	11.175	14	11	6.627
II	217	47	57	7.367	43	70	8.466
III	26	7	3	2.862	3	13	NA

(**b**)

Without Other Covariates

	hazard		logrank	
subgroup	ratio	95% C.I.	statistic	p-value
I	0.319	(0.141, 0.721)	8.364	0.0038
II	1.224	(0.809, 1.851)	0.919	0.3377
III	5.259	(1.323, 20.902)	6.808	0.0091

After Adjusting for Other Covariates

	hazard		wald	
subgroup	ratio	95% C.I.	statistic	p-value
I	0.358	(0.148, 0.865)	-2.2818	0.0230
II	0.926	(0.602, 1.424)	-0.349	0.7271
III	2.649	(0.151, 46.321)	0.667	0.5048

between a primary variable (i.e., the treatment) and other covariates. The existence of interactions is assessed by inspecting whether a nontrivial tree structure can be developed. If interactions do exist, the resultant IT structure can automatically provide a delineation of the interaction structure. In practice, it is often important to distinguish two types of interactions. If there is no directional change in terms of the comparison, the interaction is said to be *quantitative*; otherwise, it is termed *qualitative*. The presence of qualitative interactions causes much more concern than quantitative ones (see, e.g., Gail

and Simon, 1985). The results from the IT procedure allow us to address this issue. Once the interaction structure is delineated, summarizing the terminal nodes in the final tree allows for insightful exploration of the existence of possible qualitative interactions. For instance, qualitative interaction might exist among the final three subgroups in the PBC example.

The proposed IT structure helps with the identification of the most and least effective subgroups of a treatment under investigation. As demonstrated by simulation studies, one advantage of such algorithmic and adaptive methods is that they allow for empirical control of the overall type I error, which contrasts to the size issue in a designed experiment that has planned tests for well-defined subgroups. However, the IT procedure has the potential to help uncover what has not been discovered in planned subgroup analysis and generate interesting new research hypotheses. The results can be used in different ways. Take clinical trials as an example again. If the new medicine shows an overall plausible effect, and, if even in the least effective subgroup the investigational medicine does not present any harmful side effect or only the null interaction tree structure is found, then its release may be endorsed without reservation. The subgroups identified by the IT procedure may also be useful in exploring safety profiles. In trials where the proposed compound is not found to be effective, tree-structured subgroup analyses may help identify sub-populations that contribute to the failure of the compound. Information gained by using the IT procedure could be a good reference for establishing inclusion/exclusion criteria in planning future clinical trials, and as such, could be of considerable value to existing efforts to synthesize compounds for fighting deadly diseases such as cancer and HIV/AIDS.

To conclude, we emphasize the exploratory nature of subgroup analyses. There is a real danger to over-interpret the results and be over-optimistic about the findings. Some tentative guidelines for applying the IT method are in order. First, the IT method seems rather empirically conservative to the size issue. To further prevent false positive errors, we suggest applying the IT method multiple times by varying control parameters (e.g., minimum node size and maximum tree depth) and the training/test samples, especially when a non-trivial tree structure is developed via the test sample method. Secondly, there may still be considerable variations in a non-null IT structure. It might include spurious splits or omit important ones. We thus suggest that one should also consult the variable importance ranking when interpreting identified subgroups. Finally, it is a common suggestion that no conclusive inference be drawn from the results of a subgroup analysis. Same may be said for the IT procedure, the findings should never be inferred as scientific claims. Instead, they should be treated as research hypotheses to be further evaluated in future studies.

Appendix A

***Algorithm** 1: Computing Variable Importance via Random Forests.*

Initialize all V_j's to 0.
For $b = 1, 2, \ldots, B$, do

- Generate bootstrap sample $\mathcal{L}_b$ and obtain the out-of-bag sample $\mathcal{L} - \mathcal{L}_b$.
- Based on $\mathcal{L}_b$, grow a large interaction tree $\mathcal{T}_b$ by searching over m randomly selected covariates at each split.
- Send $\mathcal{L} - \mathcal{L}_b$ down T_b to compute $G(\mathcal{T}_b)$.
- For all covariates X_j, $j = 1, \ldots, p$, do
  - Permute the values of X_j in $\mathcal{L} - \mathcal{L}_b$.
  - Send the permuted $\mathcal{L} - \mathcal{L}_b$ down to $\mathcal{T}_b$ to compute $G_j(\mathcal{T}_b)$.
  - Update $V_j \leftarrow V_j + \dfrac{G(\mathcal{T}_b) - G_j(\mathcal{T}_b)}{G(\mathcal{T}_b)}$.
- End do.

End do.
Average $V_j \leftarrow V_j / B$.

References

Akaike, H. (1974). A new look at model identification. *IEEE Transactions an Automatic Control*, **19**: 716-723.

Assmann, S. F., Pocock, S. J., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis and other (mis)uses of baseline data in clinical trials. *Lancet*, **255**: 1064-1069.

Bhansali, R. J. and Downham, D. Y. (1977). Some properties of the order of an autoregressive model selected by a generalization of Akaike's FPE criterion. *Biometrika*, **64**: 547-551.

Breiman, L. (2001). Random Forests. *Machine Learning*, **45**: 5-32.

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

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

Cox, D. R. (1972). Regression models and life-tables (with discussion). *Journal of the Royal Statistical Society*, Series B, **34**: 187-202.

Cox, D. R. (1975). Partial likelihood. *Biometrika*, **62**: 269-276.

Cook, D. I., Gebski, V. J., and Keech, A. (2004). Subgroup analysis in clinical trials. *Medical Journal of Australia*, **180** (6): 289-291

Dickson, E .R., Fleming, T. R., Wiesner, R. H., Baldus, W. P., Fleming, C. R. , Ludwig, J. , and McCall, J. T. (1985). Trial of penicillamine in advanced primary biliary cirrhosis. *New England Journal of Medicine*, **312**: 1011-1015.

Fan, J. J., Su, X. G., Levine, R., Nunn, M. E., and LeBlanc, M. (2006). Multivariate Survival Trees by Goodness of Fit: Assigning Tooth Prognosis by Multivariate Survival Trees. *Journal of American Statistical Association*, **101**: 959-967.

Fleming, T. R., and Harrington, D. R. (1991). Counting processes and survival analysis. Hoboken, NJ: John Wiley & Sons:

Efron, B. (1983). Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of American Statistical Association*, **78**: 316-331.

Gail, M. and Simon, R. (1985). Testing for qualitative interactions between treatment effects and patient subsets. *Biometrics*, **41**: 362-372.

Hahn, S., Williamson, P. R., Hutton, J. L., Garner, P., and Flynn, E. V. (2000). Assessing the potential for bias in meta-analysis due to selective reporting of subgroup analyses within studies. *Statistics in Medicine*, **19**: 3325-3336.

Kalbfleisch, J. D. and Prentice, R. L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd Edition. New York: Wiley.

Keles, S. and Segal, M. (2002). Residual-based tree-structured survival analysis. *Statistics in Medicine*, **21**: 313-326.

Lagakos, S. W. (2006). The challenge of subgroup analyses - reporting without distorting. *The New England Journal of Medicine*, **354**: 1667-1669.

LeBlanc, M. and Crowley, J. (1992). Relative risk trees for censored survival data. *Biometrics*, **48**: 411-425.

Leblanc, M, Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association* **88**: 457-467.

Morgan, J. and Sonquist, J. (1963). Problems in the analysis of survey data and a proposal. *Journal of the American Statistical Association* **58**: 415-434.

Negassa, A., Ciampi, A., Abrahamowicz, M., Shapiro, S., and Boivin, J.-F., (2005). Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. *Statistics and Computing* **15**: 231-239.

Parker, A. B. and Naylor, C. D. (2000). Subgroups, treatment effects, and baseline risks: Some lessons from major cardiovascular trials. *American Heart Journal*, **139**: 952-961.

Pocock, S. J., Assmann, S. E., Enos, L. E., and Kasten, L. E. (2000). Subgroup analysis, covariate adjustment and baseline comparisons in clinical trial reporting. *Statistics in Medicine*, **21**: 2917-2930.

Sleight, P. (2000). Debate: Subgroup analyses in clinical trials fun to look at, but don't believe them! *Current Controlled Trials on Cardiovascular Medicine* **1**: 2527.

Song, Y. and Chi, G. (2007). A method for testing a prespecified subgroup in clinical trials. *Statistics in Medicine*, in press.

Su, X. G. and Fan, J. J. (2004). Multivariate survival trees: a maximum likelihood approach based on frailty models. *Biometrics*, **60**: 93-99.

van der Laan, M. J. (2006). Statistical Inference for Variable Importance. *The International Journal of Biostatistics*, Vol. **2**: Iss. 1, Article 2.