Chapter 6
The GUIDE Approach to Subgroup Identification

Wei-Yin Loh and Peigen Zhou

Abstract GUIDE is a multi-purpose algorithm for classification and regression tree construction with special capabilities for identifying subgroups with differential treatment effects. It is unique among subgroup methods in having all these features: unbiased split variable selection, approximately unbiased estimation of subgroup treatment effects, treatments with two or more levels, allowance for linear effects of prognostic variables within subgroups, and automatic handling of missing predictor variable values without imputation in piecewise-constant models. Predictor variables may be continuous, ordinal, nominal, or cyclical (such as angular measurements, hour of day, day of week, or month of year). Response variables may be univariate, multivariate, longitudinal, or right-censored. This article gives a current account of the main features of the method for subgroup identification and reviews a bootstrap method for conducting post-selection inference on the subgroup treatment effects. A data set pooled from studies of amyotrophic lateral sclerosis is used for illustration.

Keywords Bootstrap · Classification and regression tree · Confidence interval · Missing value · Post selection inference · Recursive partitioning · Variable selection

6.1 Introduction

GUIDE (Loh 2002, 2009) is an algorithm for fitting classification and regression tree models to data. AID (Morgan and Sonquist 1963) was the first algorithm but CART (Breiman et al. 1984) and RPART (Therneau and Atkinson 2018) brought the basic ideas to the mainstream. GUIDE grew out of work on an alternative approach to CART classification (Loh and Vanichsetakul 1988; Loh and Shih 1997; Kim and

W.-Y. Loh (✉) · P. Zhou
Department of Statistics, University of Wisconsin, Madison, WI, USA
e-mail: loh@stat.wisc.edu; pzhou9@wisc.edu

N. Ting et al. (eds.), *Design and Analysis of Subgroups with Biopharmaceutical Applications*, Emerging Topics in Statistics and Biostatistics,
https://doi.org/10.1007/978-3-030-40105-4_6

Loh 2001, 2003) and regression (Loh 1991b; Ahn and Loh 1994; Chaudhuri et al. 1994, 1995; Chaudhuri and Loh 2002; Chan and Loh 2004). See Loh (2014) for a recent review of classification and regression trees. Unlike AID and CART that only fit a constant in each node of the tree, GUIDE can fit linear and generalized linear models. This makes GUIDE well suited for subgroup identification—the terminal nodes of the tree are the subgroups and the regression coefficients in the node models give the treatment effects. It is unique among subgroup methods in having properties such as unbiased selection of split variables, approximately unbiased estimation of treatment effects, ability to use treatment variables with more than two levels, optional local adjustment for linear effects of prognostic variables, and automatic handling of missing values without needing prior imputation. Predictor variables may be continuous, ordinal, nominal, or cyclical (such as angles, hour of day, day of week, and month of year). Response variables may be univariate, multivariate, longitudinal, or right censored. Missing values may be coded in more than one way; for example a missing value for age of spouse may be coded as "refuse to answer" if the respondent did not provide an answer and as "valid nonresponse" if the respondent is single, widowed or divorced; see Loh et al. (2019b) for other examples.

This article gives a current account of the GUIDE method for subgroup identification. It uses data combined from several studies of ALS (Amyotrophic Lateral Sclerosis) for illustration. The data were selected because they contained all of the types of response variables that GUIDE can model and because many of the predictor variables had missing values (denoted by "NA" here). ALS is a neurological disease that affects voluntary muscle movement. Death typically occurs within 3–5 years of diagnosis. Only about a quarter of patients survive for more than 5 years after diagnosis. The data were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database (Atassi et al. 2014). In 2011, Prize4Life, in collaboration with the Northeast ALS Consortium, and with funding from the ALS Therapy Alliance, formed the PRO-ACT Consortium. The data in the PRO-ACT Database were provided by the PRO-ACT Consortium members. They were pooled from 23 completed ALS clinical trials and one observational study, and contained information on demographics, family history, and clinical and laboratory test data from more than 10700 ALS patients.

The ALS Functional Rating Scale (ALSFRS) is often used to evaluate the functional status of ALS patients. It is the sum of ten scores (speech, salivation, swallowing, handwriting, cutting food and handling utensils, dressing and hygiene, turning in bed and adjusting bed clothes, walking, climbing stairs, and breathing), with each score measured on a scale of 0–4, with 4 being normal. Seibold et al. (2016) used a subset of the data to study the effectiveness of *riluzole*, a drug approved for treatment of ALS by the US FDA, on ALSFRS at 6 months as well as survival time from trial enrollment. Using the MOB algorithm (Zeileis et al. 2008), they found that for patients with less than 468 days between disease onset and start of treatment, riluzole had a negative treatment effect on ALSFRS at 6 months.

A major difficulty with the PRO-ACT data is that besides riluzole, other medications were also tested in many of the trials (Atassi et al. 2014, Table 1).

Even worse, the additional medications were not identified in the data. To avoid confounding the effects of riluzole and that of other medications, the analysis here is restricted to the subset of 1270 subjects who were assigned to placebo or riluzole only, without other medications. Thirty-six variables were chosen as predictor variables; their names are given in Table 6.1 together with their minimum and maximum values and numbers of missing values. Three additional variables were chosen as dependent variables: (1) change in ALSFRS from baseline at 6 months, (2) monthly change in ALSFRS from baseline at months 1, 2, ..., 6, and (3) survival time in days. ALSFRS scores of subjects who had died by the time the scores were to be measured were set to 0. ALSFRS variables at 0, 1, ..., 6 months are denoted by `ALSFRS0`, `ALSFRS1`, ..., `ALSFRS6`, respectively.

6.2 Univariate Uncensored Response

Figure 6.1 shows a basic GUIDE tree for predicting change in ALSFRS after 6 months (`ALSFRS6` minus `ALSFRS0`), where a linear model (6.1) with treatment as the only predictor variable is fitted in each node. A node of the tree represents a partition of the data, with the root node corresponding to the whole data set. The sample size in each partition is printed beside the node. At each node, a variable X is selected to split the data there into two child nodes. The split, in the form $X \in A$, is printed on the left of the node. The set A is chosen to minimize the sum of the squared residuals in the left and right child nodes. Observations in the node are sent to the left child node if and only if the condition is satisfied. Node labels start with 1 for the root node; for a node with label k, its left and right child nodes are labeled $2k$ and $2k + 1$, respectively.

The root node in Fig. 6.1 is split on `Diagnosis_Delta`, which is the number of days from clinical diagnosis to the first time the subject was tested in a trial. The 239 subjects with missing values in `Diagnosis_Delta` go to terminal node 2 and the others go to intermediate node 3. It is unknown why the subjects have missing values in `Diagnosis_Delta`. One possibility is the variable was not measured in some of the trials, but this cannot be verified because trial ID was not included in the data. Nevertheless, as the barplot for node 2 in Fig. 6.1 shows, subjects in this subgroup deteriorate much more on average with riluzole than without. Subjects in node 3 are split on `Hematocrit`. Those with `Hematocrit` missing or ≤ 37.95 (abbreviated in the tree diagrams by the symbol "$\leq_*$" with the asterisk standing for "is missing") go to node 6 where they are split on `BP_Diastolic` and then on `Potassium`.

6.2.1 Node Models

Let $\mathbf{X} = (X_1, X_2, \ldots, X_K)$ denote a K-dimensional vector of covariates, Y a univariate response variable, and Z a treatment variable taking values $0, 1, \ldots, G$,

Table 6.1 Predictor variables, minimum and maximum values, numbers of categorical levels, and numbers of missing values for modeling the difference `ALSFRS6-ALSFRS0`

Name	Definition	Min	Max	Miss
Demographics_Delta	Demographic measurement day	−35.00	32.00	19
Age	Subject age at start of trial	18.00	82.00	
Sex	Subject gender (female, male)			
Race	Subject race (5 categories)			3
ALS_History_Delta	Day ALS history reported	0.00	3.00	43
Symptom	Major symptom (10 categories)			1085
Onset_Delta	Day of disease onset, from first test	−1900.00	−84.00	47
Diagnosis_Delta	Day of diagnosis, from first test	−1666.00	0.00	239
Site_of_Onset	Site of disease onset (3 categories)			
Albumin	Albumin in blood (g/L)	31.67	53.00	332
ALT_SGPT	Alanine amino transferase (U/L)	6.00	181.00	259
AST_SGOT	Aspartate amino transferase (U/L)	7.50	116.00	258
Basophil_Count	Amount in white blood cell ($\times 109$/L)	0.00	5.56	341
Basophils	Percent in white blood cell count	0.00	3.00	365
Blood_Urea_Nitrogen	Ureas (mmol/L)	0.95	17.34	218
Calcium	Calcium in metabolic panel (mmol/L)	1.55	3.00	333
Creatinine	Creatinine from kidney test	25.00	159.10	216
Eosinophils	Percent in white blood cell count	0.00	15.00	365
Glucose	Glucose in blood (mmol/L)	0.07	18.56	325
Hematocrit	Percent red blood cells	0.00	56.00	326
Hemoglobin	Hemoglobin in blood (g/L)	94.50	181.00	326
Lymphocytes	Percent lymphocyte in blood	8.70	50.00	365
Monocytes	Percent in white blood cell count	0.00	21.40	365
Platelets	Platelets in blood ($\times 109$/L)	0.20	552.00	332
Potassium	Potassium in electrolytes (mmol/L)	3.30	5.50	258
Sodium	Sodium in electrolytes (mmol/L)	125.00	150.00	257
Urine_Ph	Acidity of urine	5.00	9.00	355
SVC (Slow_vital_Capacity)	Volume of air exhaled slowly (L)	1.00	7.00	737
Slow_vital_Capacity_ Delta	Day of SVC assessment	0.00	14.00	737
BP_Diastolic	Diastolic blood pressure (mmHg)	52.00	125.00	217
BP_Systolic	Systolic blood pressure (mmHg)	90.00	200.00	217
Height	Subject height (in)	131.00	205.00	225
Pulse	Beats per minute	42.00	120.00	218
Weight	Subject weight (kg)	38.33	138.20	178
ALSFRS0	ALSFRS at baseline	10.00	40.00	
ALSFRS_Delta0	Day of ALSFRS0 measurement	−7.00	154.00	

Variables with names containing "_Delta" are days from trial onset to the date that an assessment took place, with negative values for occurrences before trial onset

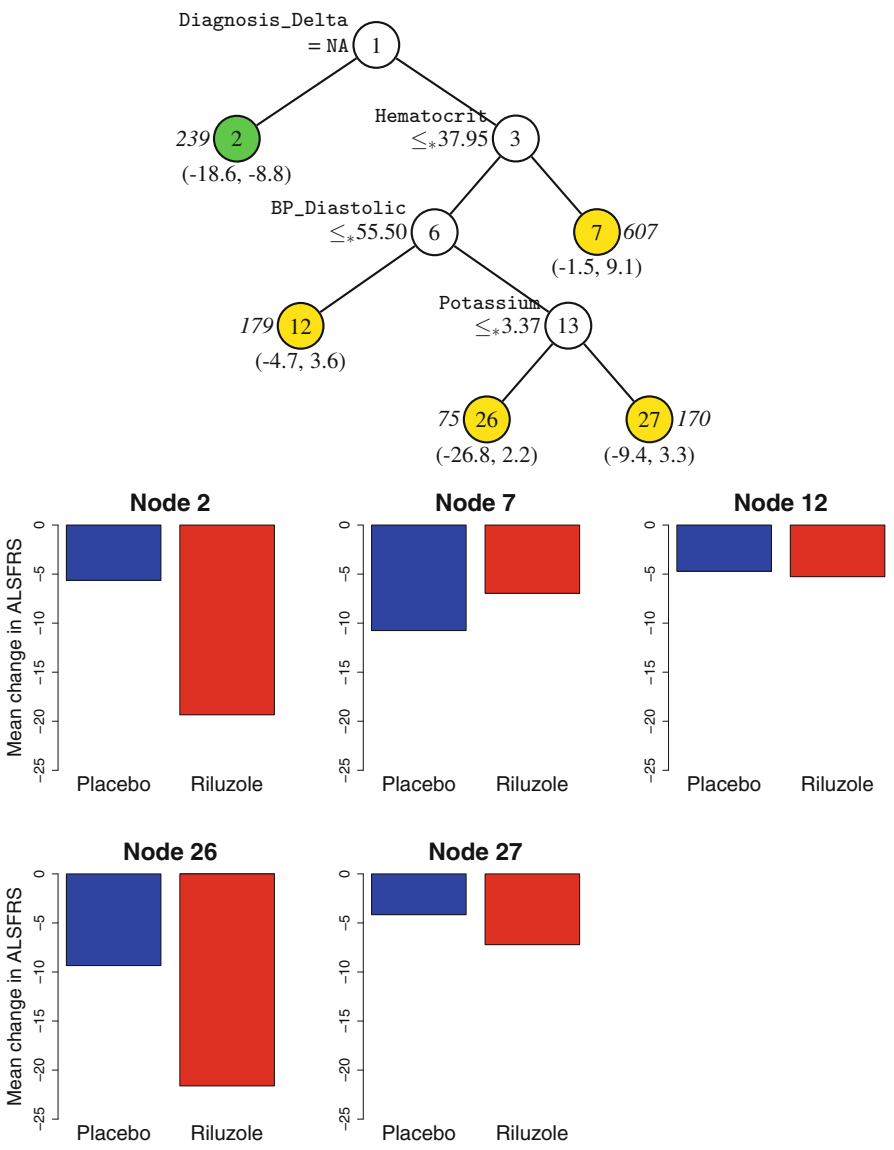

Fig. 6.1 GUIDE tree for change in ALSFRS (`ALSFRS6-ALSFRS0`) using 1270 observations and node model (6.1). At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol "$\leq_*$" stands for "$\leq$ or missing." Sample sizes *(in italics)* are printed beside nodes. Bootstrap-calibrated 90% simultaneous intervals of treatment effect are given below nodes. Calibrated alpha is 1.3×10^{-5}. Treatment effect is statistically significant in green node. Barplots show means of change in ALSFRS for placebo and riluzole subjects in the terminal nodes

with 0 being the reference (or placebo) level. Let t denote a node of the tree. A regression tree model is constructed by recursively partitioning a training sample into subsets that are represented by the nodes of a tree. A large majority of regression tree methods for subgroup identification employ stopping rules based on Bonferroni-corrected p-values (Lipkovich et al. 2011; Seibold et al. 2016; Su et al. 2009). Other methods (Dusseldorp and Meulman 2004; Foster et al. 2011), including GUIDE, first grow an overly large tree and then use cross-validation to prune it to a smaller size. We only describe the GUIDE node fitting and splitting steps here because the pruning step is the same as that of CART.

For least-squares regression, GUIDE fits a linear model $Y = f(\mathbf{X}, Z) + \epsilon$ to the data in each node of a tree; ϵ is an independent zero-mean random variable with variance that is constant within each node but may vary between nodes. Four choices of $f(\mathbf{x}, z)$ are available, depending on the number of X variables to be included. Let β_z $(z = 1, 2, \ldots, G)$ denote the effect of treatment level z (versus level 0). The choices are:

$$f(\mathbf{x}, z) = \eta + \beta_z \quad \text{(Treatment only)} \tag{6.1}$$

$$f(\mathbf{x}, z) = \eta + \beta_z + \sum_{j=1}^{p} \gamma_j x_{k^*}^j \quad \text{(Polynomial of degree } p\text{)} \tag{6.2}$$

$$f(\mathbf{x}, z) = \eta + \beta_z + \sum_{k}^{K} \gamma_k x_k \quad \text{(Multiple linear)} \tag{6.3}$$

$$f(\mathbf{x}, z) = \eta + \beta_z + \sum_{k \in S} \gamma_k x_k \quad \text{(Stepwise linear)} \tag{6.4}$$

In (6.2), p is a user-specified positive integer and k^* is the value of k such that X_k minimizes the sum of squared residuals in the node (k^* may vary from node to node). In (6.4), the set S is the set of indices of the variables X_k that are selected by forward and backward stepwise regression in the node. Thus the model for a tree with terminal nodes $t_1, t_2, \ldots, t_\tau$ may be written as

$$Y = \begin{cases} f_1(\mathbf{X}, Z) + \epsilon_1, \ \mathbf{X} \in t_1 \\ \quad \vdots \\ f_\tau(\mathbf{X}, Z) + \epsilon_\tau, \ \mathbf{X} \in t_\tau \end{cases} \tag{6.5}$$

where $f_1, f_2, \ldots, f_\tau$ take one of the functional forms (6.1)–(6.4) and $\epsilon_1, \ldots, \epsilon_\tau$ are independent random variables with mean zero and variances $\sigma_1^2, \ldots, \sigma_\tau^2$. This is different from the model

$$Y = \sum_{j=1}^{\tau} f_j(\mathbf{X}, Z)\, I(\mathbf{X} \in t_j) + \epsilon \tag{6.6}$$

which assumes that the error variance is the same in all nodes. The least-squares estimates of the regression coefficients are the same in models (6.5) and (6.6), but not their standard error estimates. In (6.2)–(6.4), missing values in the X variables are imputed by their node means.

Figure 6.1 was constructed using model (6.1) and Fig. 6.2 was constructed using model (6.2) with $p = 1$. The name of the best linear prognostic variable X_{k^*} is given beneath each terminal node. The root node splits on "`Diagnosis_Delta` ≤ -1072 or missing." Of the 245 subjects in this subgroup, 239 are missing `Diagnosis_Delta`. The best linear prognostic variable in node 2 is `Pulse`. Plots of the data and regression lines for placebo and riluzole subjects in each node are shown in the lower half of Fig. 6.2. Mean imputation of `Sodium` is clearly shown by the vertical line of points in the plot of node 13.

6.2.2 Split Variable Selection

To find a variable to split a node t, a test of treatment-covariate interaction is performed for each X_k on the data in t. (This is the default "Gi" method.) Let n_t denote the number of observations in t. The following steps are carried out for each variable X_j, $j = 1, 2, \ldots, K$.

1. If X_j is a categorical variable, define $V = X_j$ and let h denote its number of levels (including a level for `NA`, if any).
2. If X_j is ordinal and takes only one value (including `NA`) in the node, do not use it to split the node. Otherwise, let m denote the number of distinct values (including `NA`) of X_j in t. Transform it to a discrete variable V with h values as follows.
   (a) If $m \leq 4$ or if $m = 5$ *and* X_j has missing values, define $h = m$. Otherwise, define $h = 3$ if $n_t < 30(G + 1)$ and $h = 4$ otherwise.
      (i) If X_j has missing values in t, define $r_k = k/(h-1), k = 1, 2, \ldots, h-2$.
      (ii) If X_j has no missing values in t, define $r_k = k/h, k = 1, 2, \ldots, h-1$.
   (b) Define $q_0 = -\infty$ and let q_k $(k > 0)$ be the sample r_k-quantile of X_j in t.
      (i) If X_j has missing values in t, define

$$V = \sum_{k=1}^{h-2} kI(q_{k-1} < X_j \leq q_k) + (h-1)I(X_j > q_{h-2}) + hI(X_j = \texttt{NA}).$$

      (ii) If X_j has no missing values in t, define

$$V = \sum_{k=1}^{h-1} kI(q_{k-1} < X_j \leq q_k) + hI(X_j > q_{h-1}).$$

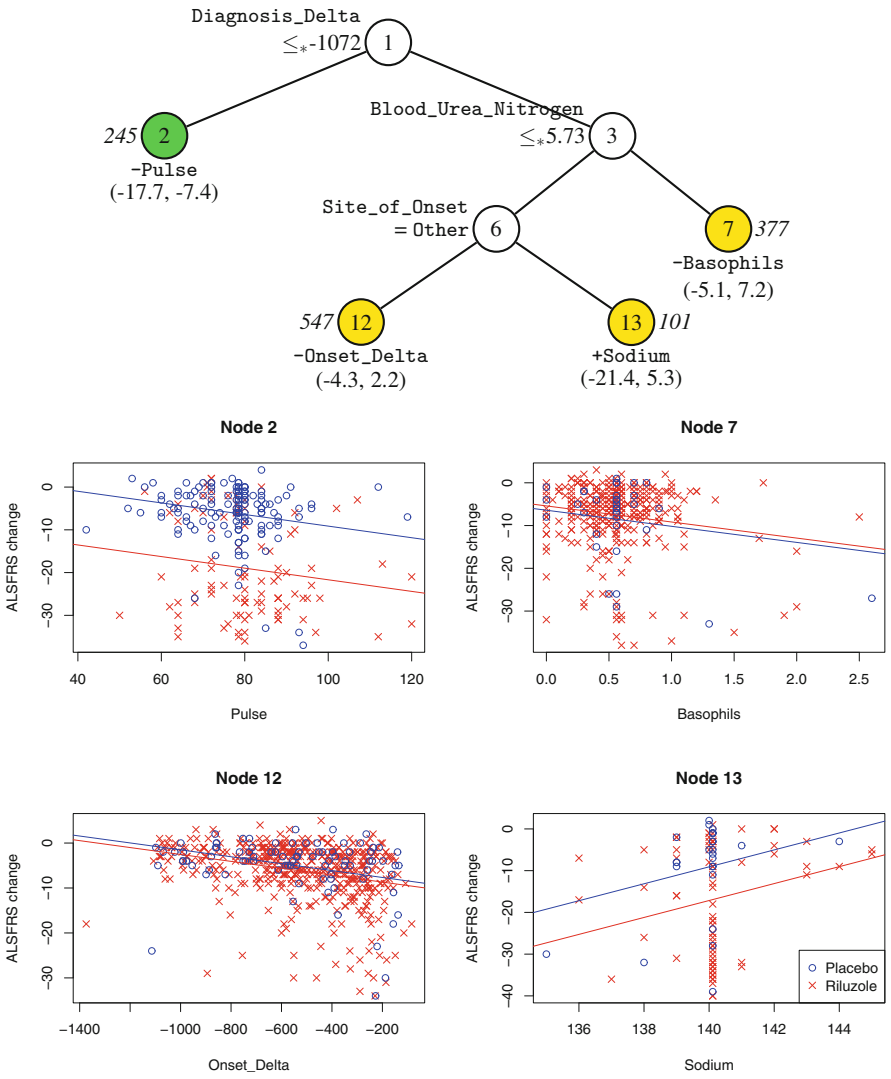

Fig. 6.2 GUIDE tree for `ALSFRS6-ALSFRS0` using 1270 observations and node model (6.2) with polynomials of degree 1. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Sample sizes (*in italics*) are printed beside nodes. Name of best linear prognostic variable (with sign of slope) and bootstrap-calibrated 90% simultaneous confidence interval for treatment effect are below each node. Calibrated alpha is 8.9×10^{-6}. Treatment effect is statistically significant in green node. Plots of change in ALSFRS versus best linear predictor show data points and fitted regression lines in the terminal nodes. Missing values in predictor variables are imputed by the means of their non-missing values in the nodes

3. Test the additive model $E(Y|Z, V) = \eta + \sum_z \beta_z I(Z = z) + \sum_v \gamma_v I(V = v)$, with $\beta_0 = \gamma_1 = 0$, against the full model $E(Y|Z, V) = \sum_z \sum_v \omega_{vz} I(V = v, Z = z)$ and obtain the p-value p_j.

Split node t on the X_j with the smallest value of p_j.

6.2.3 Split Set Selection

After X is selected, a search is carried out for the best split "$X \in A$", where A depends on whether X is ordinal or categorical.

6.2.3.1 Ordinal Variable

If X is ordinal, three types of splits are evaluated.

1. $X =$ `NA`: an observation goes to the left node if and only if its value is missing.
2. $X =$ `NA` or $X \leq c$: an observation goes to the left node if and only if its value is missing or if it is less than or equal to c.
3. $X \leq c$: an observation goes to the left node if and only if its value is not missing and it is less than or equal to c.

Candidate values of c are the midpoints between consecutive order statistics of X in t. If X has m order statistics, the maximum number of possible splits is $(m - 1)$ or $\{1 + 2(m - 1)\}$, depending on the absence or presence of missing X values in t. Permissible splits are those that yield two child nodes with each having two or more observations per treatment. The selected split is the one that minimizes the sum of the deviances (or sum of squared residuals in the case of least-squares regression) in the two child nodes.

This method of dealing with missing values is unique to GUIDE. CART uses a hierarchical system of "surrogate splits" on alternative X variables to send observations with missing values to the child nodes. Because the surrogate splits depend on the (missing and non-missing) values of the alternative X variables, observations with missing values do not necessarily go to the same child node. Therefore it is impossible to predict the path of an observation by looking at the tree without knowing the values of its predictor variables. Besides, CART's surrogates splits are biased towards X variables with few missing values (Kim and Loh 2001). Other subgroup methods are typically inapplicable to data with missing values (Dusseldorp and Meulman 2004; Su et al. 2009; Foster et al. 2011; Seibold et al. 2016).

Sometimes, missing-value imputation is illogical, e.g., a prostate-specific antigen test result for a female subject or the age of first cigarette for a subject who never smoked. Other times, imputation erases useful information. For example, if missing values were imputed before application of GUIDE, the large difference in treatment effect between subjects with and without missing values in `Diagnosis_Delta` in Fig. 6.1 would be undetected.

6.2.3.2 Categorical Variable

If X is a categorical variable, the split has the form $X \in A$, where A is a non-trivial subset of the values (including NA) of X in t. A complete search of all possible values of A can be computationally expensive if the number, m, of distinct values (including NA) of X in t is large, because there are potentially $(2^{m-1}-1)$ splits (less if some splits yield child nodes with fewer than two observations per treatment). Therefore GUIDE carries out a complete search only if $m \leq 11$. If $m > 11$, it performs an approximate search by means of linear discriminant analysis , based on an idea from Loh and Vanichsetakul (1988), Loh and Shih (1997), and Loh (2009).

1. Let $\bar{y}_z$ denote the sample mean of the Y values in t that belong to treatment $Z = z$ $(z = 0, 1, \ldots, G)$.
2. Define the class variable

$$C = \begin{cases} 2z - 1, & \text{if } Z = z \text{ and } Y > \bar{y}_z \\ 2z, & \text{if } Z = z \text{ and } Y \leq \bar{y}_z. \end{cases}$$

3. Let $\{a_1, a_2, \ldots, a_m\}$ denote the categorical values of X in t. Transform X to an m-dimensional 0–1 dummy vector $\mathbf{D} = (D_1, D_2, \ldots, D_m)$, where $D_i = I(X = a_i)$, $i = 1, 2, \ldots, m$.
4. Apply linear discriminant analysis to the data $(\mathbf{D}, C)$ in t to find the discriminant variables $B_j = \sum_{i=1}^{m} b_{ij} D_i$, $j = 1, 2, \ldots$. These variables are also called canonical variates (Gnanadesikan 1997).
5. For each j, find the split $B_j \leq c_j$ that minimizes the sum of the squared residuals of the least-squares models fitted in the child nodes induced by the split.
6. Let j^* be the value of j for which $B_j \leq c_j$ has a smallest sum of squared residuals.
7. Split the node with $B_{j^*} \leq c_{j^*}$. Because $B_{j^*} = \sum_{i=1}^{m} b_{ij^*} D_i = \sum_{i=1}^{m} b_{ij^*} I(X = a_i)$, the split is equivalent to $X \in A$ with $A = \{a_i : b_{ij^*} \leq c_{j^*}\}$.

6.3 Bootstrap Confidence Intervals

The barplots in the lower half of Fig. 6.1 show that the subgroups defined by nodes 2 and 26 have the largest treatment effects. Similarly, the graphs in the lower half of Fig. 6.2 suggest that node 2 has the largest treatment effect. Are the effects statistically significant? This question cannot be answered by means of traditional methods because the subgroups were not specified independently of the data. It is a question of *post-selection inference*.

Given node t and z, let $\hat{\beta}(t, z)$ be the estimated treatment effect for $Z = z$ in t, let $\hat{\sigma}_\beta(t, z)$ denote its usual estimated standard error, and let ν_t be the residual degrees of freedom. Further, let $t_{\nu,\alpha}$ denote the $(1-\alpha)$-quantile of the t-distribution with ν degrees of freedom and let τ denote the number of terminal nodes of the tree. Let

Table 6.2 90% simultaneous intervals for subgroup treatment effects in Figs. 6.1 and 6.2

Model	Node	$B(0.10, t, z)$	$J(\alpha_{\hat{F}}, t, z)$
Figure 6.1	2	$(-16.0, -11.4)$	$(-18.7, -8.7)$
$\alpha_{\hat{F}} = 1.3 \times 10^{-5}$	7	$(1.3, 6.3)$	$(-1.6, 9.2)$
	12	$(-2.5, 1.4)$	$(-4.7, 3.6)$
	26	$(-18.7, -5.8)$	$(-26.2, 1.7)$
	27	$(-6.0, -0.1)$	$(-9.4, 3.3)$
Figure 6.2	2	$(-15.1, -10.0)$	$(-17.7, -7.4)$
$\alpha_{\hat{F}} = 8.9 \times 10^{-6}$	7	$(-2.1, 4.1)$	$(-5.1, 7.2)$
	12	$(-2.7, 0.6)$	$(-4.3, 2.2)$
	13	$(-14.5, -1.6)$	$(-21.4, 5.3)$

$$B(\alpha, t, z) = \hat{\beta}(t, z) \pm t_{\nu_t, \alpha/(2\tau)}\, \hat{\sigma}_\beta(t, z) \tag{6.7}$$

be the Bonferroni-corrected $100(1-\alpha)\%$ simultaneous t-interval for the treatment effect of $Z = z$ in node t. The middle column of Table 6.2 gives the values of $B(0.10, t, z)$ for the trees in Figs. 6.1 and 6.2. Despite the Bonferroni correction, the standard errors $\hat{\sigma}_\beta(t, z)$ are biased low because they do not account for the uncertainty due to split selection. As a result, the intervals $B(\alpha, t, z)$ tend to be too short and their simultaneous coverage probability is less than $(1-\alpha)$.

There are two obvious ways to lengthen the interval widths to improve their coverage probabilities. One is to correct the standard error estimates, but this is formidable due to the complexity of the tree algorithm. Another way is to reduce the nominal value of α in (6.7). For example, to obtain 90% simultaneous coverage, we could use $B(\alpha, t, z)$ with a nominal $\alpha < 0.10$. To find the right nominal value of α, we first need to define the *estimand* of $\hat{\beta}(t, z)$, which is the true treatment effect in t. Let $\hat{F}$ denote the training data and F the population from which they are drawn. By definition, $\hat{\beta}(t, z)$ $(z = 1, \ldots, G)$ are the values of the treatment effect coefficients that minimize $\sum_{i \in t}(y_i - f(\mathbf{x}_i, z_i))^2$, where the sum is over the observations in node t. Their estimands, denoted by are $\beta_F(t, z)$, are the values of the treatment effect coefficients that minimize $E\{(Y - f(\mathbf{X}, Z))^2 I(\mathbf{X} \in t)\}$. Clearly, $\beta_F(t, z)$ is a random variable, because it depends on t, which in turn depends on $\hat{F}$. If F is known and t is given, however, $\beta_F(t, z)$ can be computed, by simulation from F if necessary.

Let $J(\alpha, t, z) = \hat{\beta}(t, z) \pm t_{\nu_t, \alpha/2}\, \hat{\sigma}_\beta(t, z)$ denote the nominal $100(1-\alpha)\%$ t-interval, let $\tilde{T}$ be the set of terminal nodes, and let $\gamma_F(\alpha) = P[\cap_{t \in \tilde{T}} \{\beta_F(t, z) \in J(\alpha, t, z)\}]$ denote the simultaneous coverage probability. Clearly, $\gamma_F(\alpha) \uparrow 1$ as $\alpha \downarrow 0$. Given a desired simultaneous coverage probability γ^*, let α_F be the solution of the equation $\gamma_F(\alpha_F) = \gamma^*$. Then the intervals $J(\alpha_F, t, z)$ have exact simultaneous coverage γ^*. We call α_F the "calibrated α." Note that there is no need to work with the Bonferroni-corrected interval (6.7) because $\gamma_F(\alpha)$ is, by definition, a simultaneous coverage probability.

Of course, the value of α_F is not computable if F is unknown. In that case, a natural solution is *bootstrap calibration*, a method proposed in Loh (1987, 1991a) for the simpler problem of estimating a population mean. It was extended to

Algorithm 1: Bootstrap calibration of confidence intervals for treatment effects

Data: Given $K > 0$ and $\alpha \in (0, 1)$, $\alpha_1 < \alpha_2 < \ldots < \alpha_K = \alpha$; tree T with nodes $t_1, t_2, \ldots, t_L$ constructed from $\mathscr{D} = \{(\mathbf{X}_i, Y_i, Z_i), i = 1, 2, \ldots, n\}$; and model M (one of (6.1), ..., or (6.4)) based on T with estimated treatment effects $\hat{\beta}_{tz}$, $z = 1, 2, \ldots, G$; $t = t_1, t_2, \ldots, t_L$.
Result: $(1 - \alpha)$ simultaneous t-intervals for $\{\beta_{tz}\}$.
begin
  $\gamma_k \leftarrow 0$ for $k = 1, 2, \ldots, K$;
  for $b \leftarrow 1$ **to** B **do**
    bootstrap data $\mathscr{D}_b^* = \{(\mathbf{X}_i^*, Y_i^*, Z_i^*), i = 1, 2, \ldots, n\}$ from $\mathscr{D}$;
    construct from $\mathscr{D}_b^*$ tree T_b with nodes $t_{b1}^*, t_{b2}^*, \ldots, t_{bL_b}^*$;
    fit model M based on T_b to $\mathscr{D}$ observations to get "true" effects $\beta(t_{bl}^*, z)$; $z = 1, \ldots, G$; $l = 1, \ldots, L_b$;
    fit model M based on T_b to $\mathscr{D}_b^*$ observations to get estimates $\hat{\beta}(t_{bl}^*, z)$, residual degrees of freedom ν_{bl} and standard errors $\hat{\sigma}_\beta(t_{bl}^*, z)$; $z = 1, \ldots, G$; $l = 1, \ldots, L_b$;
    for $z \leftarrow 1$ **to** G **do**
      for $l \leftarrow 1$ **to** L_b **do**
        for $k \leftarrow 1$ **to** K **do**
          $J_{klz} \leftarrow (1 - \alpha_k)$ t-interval $\hat{\beta}(t_{bl}^*, z) \pm t_{\nu_{bl}, \alpha_k/2} \hat{\sigma}_\beta(t_{bl}^*, z)$;
          if $\beta(t_{bl}^*, z) \in J_{klz}$ **then**
            $c_{klz} \leftarrow 1$; `/* interval contains true beta */`
          else
            $c_{klz} \leftarrow 0$; `/* interval does not contain true beta */`
          end
        end
      end
    end
    for $k \leftarrow 1$ **to** K **do**
      if $\min_{lz} c_{klz} = 1$ **then**
        $\gamma_k \leftarrow \gamma_k + 1$
      end
    end
  end
  $\gamma_k \leftarrow \gamma_k / B$ for $k = 1, 2, \ldots, K$;
  $q \leftarrow$ smallest k such that $\gamma_k < 1 - \alpha$;
  $g \leftarrow (\gamma_{q-1} - 1 + \alpha)/(\gamma_{q-1} - \gamma_q)$;
  $\alpha' \leftarrow (1 - g)\alpha_{q-1} + g\alpha_q$;
  construct $(1 - \alpha')$ simultaneous t-intervals for β_{tz} for $t = t_1, t_2, \ldots, t_L$; $z = 1, \ldots, G$
end

estimation of subgroup treatment effects in Loh et al. (2016, 2019c). The idea is to replace F with $\hat{F}$ in the calculations. Specifically, use simulation from $\hat{F}$ to find the solution $\alpha_{\hat{F}}$ of the equation $\gamma_{\hat{F}}(\alpha_{\hat{F}}) = \gamma^*$. The resulting intervals $J(\alpha_{\hat{F}}, t, z)$ are called "bootstrap-calibrated" $100\gamma^*\%$ simultaneous intervals. Algorithm 1 gives the instructions in pseudo-code, using a grid search to find $\alpha_{\hat{F}}$. The numerical results here (including those in the last column of Table 6.2) were obtained with a grid of 200 nominal values of α and 1000 bootstrap iterations. Simultaneous

90% bootstrap-calibrated intervals of treatment effect are given beneath the terminal nodes of the trees in Figs. 6.1 and 6.2. Their respective bootstrap-calibrated alpha values are $\alpha_{\hat{F}} = 1.3 \times 10^{-5}$ and 8.9×10^{-6}. In the tree diagrams, nodes with statistically significant treatment effects are in green color.

6.4 Multivariate Uncensored Responses

GUIDE can construct a least-squares regression tree for data with longitudinal or multivariate response variables as well. Given d response variables $Y_1, Y_2, \ldots, Y_d$, it fits the treatment-only model $E(Y_j|Z) = \eta_j + \sum_{z=1}^{G} \beta_{jz} I(Z = z)$, $j = 1, \ldots, d$, separately to each variable in each node. To find the variable to split a node, the test for treatment-covariate interaction in Sect. 6.2.2 is performed d times for each X_i (once for each Y_j) to obtain the p-value $p_{i1}, p_{i2}, \ldots, p_{id}$. Let $\chi^2_{\nu,\alpha}$ denote the $(1-\alpha)$-quantile of the chi-squared distribution with ν degrees of freedom. The variable X_i for which $\sum_{j=1}^{d} \chi^2_{1,p_{ij}}$ is maximum is selected to split the node. To allow for correlations in the response variables, GUIDE can optionally apply the treatment-covariate interaction tests to the principal component (PC) or linear discriminant (LD) variates computed from the Y_j values in the node. Specifically, if principal component transformation is desired, the $(Y_1, Y_2, \ldots, Y_d)$ data vectors in the node are transformed to their PCs $(Y'_1, Y'_2, \ldots, Y'_d)$ first; then the treatment-covariate interactions tests are applied to the $(Y'_1, Y'_2, \ldots, Y'_d)$ data vectors. Similarly, if LD is desired, the $(Y_1, Y_2, \ldots, Y_d)$ data vectors in the node are transformed to their linear discriminant variates, using the treatment levels as class labels. The PC and LD transformations are carried out *locally* at each node. After the split variable X_i is selected, its split point (if X_i is ordinal) or split set (if X_i is categorical) is the value that yields the smallest total sum of squared residuals (where the total is over the d models $E(Y_j|Z) = \eta_j + \sum_z \beta_{jz} I(Z = z)$) in the left and right child nodes. See Loh and Zheng (2013) and Loh et al. (2016) for more details.

Using change from baseline of `ALSFRS1`, `ALSFRS2`, ..., `ALSFRS6` as longitudinal response variables, only the PC option yielded a nontrivial tree, as shown in Fig. 6.3. Subjects who died after 6 months and had missing values in any response variable were omitted, leaving a training sample of 627 observations. The tree has only one split, the same as the split at the root node of Fig. 6.2. The plots below the tree diagram show bootstrap-calibrated 90% simultaneous intervals for the treatment effect for each response variable in each terminal node. The longer lengths of the intervals in the left node are due to its much smaller sample size. Because every interval contains 0, there is no subgroup with statistically significant treatment effect.

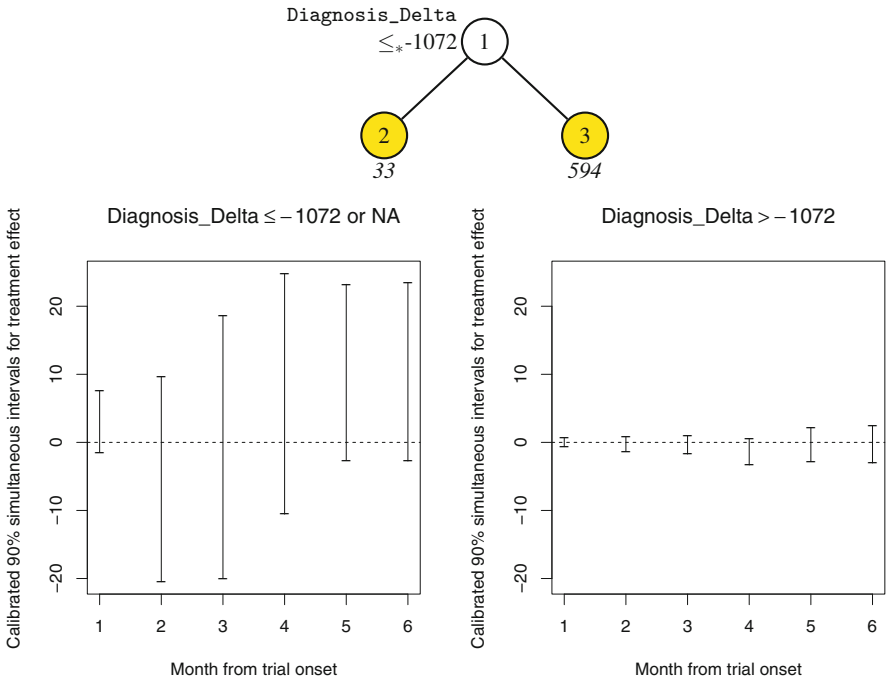

Fig. 6.3 GUIDE tree for change from baseline of longitudinal responses `ALSFRS1`, `ALSFRS2`, ..., `ALSFRS6`, using 627 observations and PCA at each node. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Sample size *(in italics)* printed below nodes. Bootstrap-calibrated 90% simultaneous intervals for treatment effect of each response variable in each node plotted below tree. Calibrated alpha is 0.011

6.5 Time-to-Event Response

Let $(U_1, \mathbf{X}_1)$, $(U_2, \mathbf{X}_2)$, ..., $(U_n, \mathbf{X}_n)$ be the survival times and predictor variable values of n subjects. Let $V_1, V_2, \ldots, V_n$ be independent and identically distributed observations from a censoring distribution and let $\delta_i = I(U_i < V_i)$ be the event indicator. The observed data vector of subject i is $(Y_i, \delta_i, \mathbf{X}_i)$, where $Y_i = \min(U_i, V_i)$. Let $\lambda(y, \mathbf{x}, z)$ denote the hazard function at time y and covariates $\mathbf{x}$ and z. The proportional hazards model stipulates that $\lambda(y, \mathbf{x}, z) = \lambda_0(y)\exp(\eta)$, where $\lambda_0(y)$ is a baseline hazard function independent of $(\mathbf{x}, z)$, and η is a function of $\mathbf{x}$ and z. Many methods fit a proportional hazards model to the data in each node separately (Negassa et al. 2005; Su et al. 2009; Lipkovich et al. 2011; Lipkovich and Dmitrienko 2014; Seibold et al. 2016), giving the tree model

$$\lambda(y, \mathbf{x}, z) = \sum_{j=1}^{\tau} \lambda_{j0}(y)\exp(\eta_j + \beta_{jz})I(\mathbf{x} \in t_j);\ \ \beta_{j0} = 0;\ j = 1, 2, \ldots, \tau.$$

Because the baseline hazard $\lambda_{j0}(y)$ varies from node to node, the model does not have proportional hazards overall. Therefore estimates of regression coefficients cannot be compared between nodes and relative risks are not independent of y.

GUIDE (Loh et al. 2015) fits one of the following three truly proportional hazards models instead.

$$\lambda(y, \mathbf{x}, z) = \lambda_0(y) \exp\left[\sum_{j=1}^{\tau} \{\eta_j + \beta_{jz}\} I(\mathbf{x} \in t_j)\right] \tag{6.8}$$

$$\lambda(y, \mathbf{x}, z) = \lambda_0(y) \exp\left[\sum_{j=1}^{\tau} \left\{\eta_j + \beta_{jz} + \sum_{i=1}^{p} \gamma_{ji} x_{k*}^i\right\} I(\mathbf{x} \in t_j)\right] \tag{6.9}$$

$$\lambda(y, \mathbf{x}, z) = \lambda_0(y) \exp\left[\sum_{j=1}^{\tau} \left\{\eta_j + \beta_{jz} + \sum_{k}^{K} \delta_{jk} x_k\right\} I(\mathbf{x} \in t_j)\right] \tag{6.10}$$

where $\beta_{j0} = 0$ $(j = 1, \ldots, \tau)$ and the η_j satisfy a constraint such as $\sum_j \eta_j = 0$ to prevent over-parameterization. Model fitting is carried out by means of a well-known connection between proportional hazards regression and Poisson regression (Aitkin and Clayton 1980; Laird and Olivier 1981). Let $\Lambda_0(y) = \int_{-\infty}^{y} \lambda_0(u)\, du$ denote the baseline cumulative hazard function. The regression coefficients in (6.8), (6.9), or (6.10) are estimated by iteratively fitting a GUIDE Poisson regression tree (Chaudhuri et al. 1995; Loh 2006), using the event indicators δ_i as Poisson responses, $\log \Lambda_0(y_i)$ as offset variable, and the Poisson models

$$\log E(\delta|Z) = \log \Lambda_0(y) + \xi_j + \sum_z \beta_{jz} I(Z = z),$$

$$\log E(\delta|Z, X_{k*}) = \log \Lambda_0(y) + \xi_j + \sum_z \beta_{jz} I(Z = z) + \sum_{i=1}^{p} \gamma_{ji} X_{k*}^i,$$

$$\log E(\delta|Z, X_1, X_2, \ldots, X_k) = \log \Lambda_0(y) + \xi_j + \sum_z \beta_{jz} I(Z = z) + \sum_{k}^{K} \delta_{jk} X_k,$$

respectively, in each node t_j. At the first iteration, $\Lambda_0(y_i)$ is estimated by the Nelson-Aalen method (Aalen 1978; Breslow 1972). Then the estimated relative risks of the observations from the tree model are used to update $\Lambda_0(y_i)$ for the next iteration; see, e.g., Lawless (1982, p. 361).

Figure 6.4 gives the result of fitting model (6.8) from the 966 subjects with non-missing censored or observed survival time in the ALS data. The tree splits on `Symptom` to give two terminal nodes. The left node consists of 815 subjects with `Symptom` either missing or is `speech`. The other 151 subjects go to the right node, which has a statistically significant treatment effect based on the bootstrap-

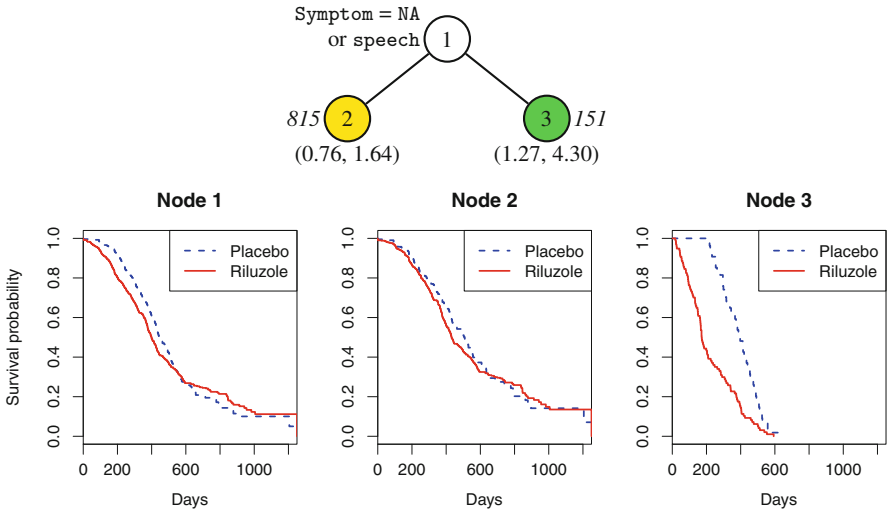

Fig. 6.4 GUIDE proportional hazards regression tree for differential treatment effects using model (6.8). Kaplan-Meier survival curves in each node are shown below the tree. Numbers in italics beside terminal nodes are sample sizes. Bootstrap-calibrated 90% simultaneous confidence intervals of relative risks (riluzole versus placebo) are given below terminal nodes. Calibrated alpha is 0.0003. Treatment effect is statistically significant in green node

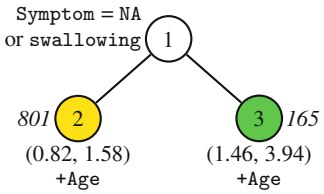

Fig. 6.5 GUIDE proportional hazards regression tree for differential treatment effects using model (6.9). Numbers in italics beside terminal nodes are sample sizes. Bootstrap-calibrated 90% simultaneous confidence intervals of relative risks (riluzole versus placebo) and name of linear prognostic variable (with sign indicating slope) are given below nodes. Calibrated alpha is 0.003. Treatment effect is statistically significant in green node

calibrated 90% simultaneous confidence intervals of relative risks printed below the nodes. Kaplan-Meier survival curves for placebo and riluzole subjects in each node are shown below the tree diagram.

Figure 6.5 gives the result for model (6.9) with polynomial degree $p = 1$. The root node is split into two terminal nodes on `Symptom`, but now the model in each node includes the effect of the best linear prognostic variable (which turns out to be `Age` in both child nodes). According to the bootstrap-calibrated 90% simultaneous intervals for relative risk printed below the nodes, the subgroup with significant treatment effect consists of subjects for which `Symptom` is neither missing nor `swallowing`.

6.6 Concluding Remarks

We have explained and demonstrated the main features of the GUIDE method for subgroup identification and discussed a bootstrap method of confidence interval construction for subgroup treatment effects. The bootstrap method is quite general and is applicable to algorithms other than GUIDE. Because it expands the traditional t-intervals to account for uncertainty due to split selection, it is more efficient if the estimated subgroup treatment effects are unbiased. The method may still be applicable if the estimates are biased, but the calibrated intervals would be wider as a result. Biased estimates of subgroup treatment effects are common among algorithms that search for splits to maximize the difference in treatment effects in the child nodes. A comparison of methods on this and other criteria is reported in a forthcoming article (Loh et al. 2019a).

Although GUIDE does not impute missing values for split selection, it does impute them in the predictor variables with their node sample means when fitting models (6.2)–(6.4) in the nodes. Therefore these models, e.g., Figs. 6.2 and 6.5, assume that missing values in the X variables are missing at random (MAR). But the MAR assumption is not needed for model (6.1), such as Figs. 6.1, 6.3, and 6.4.

There are two newer GUIDE features that are not discussed here. One is cyclic or periodic predictor variables, such as angle of impact in an automobile crash, day of week of hospital admission, and time of day of medication administration. If GUIDE splits a node on such a variable, the split takes the form of a finite interval of values $a < X \leq b$ instead of a half-line $X \leq c$. Another feature is accommodation of multiple missing-value codes. For example, the result of a lab test may be "missing" for various reasons. It may not have been ordered by the physician because it was risky for the patient, it may be inappropriate (e.g., a mammogram for a male or a prostate-specific antigen test for a female), the patient may have declined the test due to cost, or the result of the test was accidentally or erroneously not reported. If the "missing" values are all recorded as `NA`, a split would take the form "$X \leq c$ or $X =$ `NA`" or "$X \leq c$ and $X \neq$ `NA`". But if the reasons for missingness are known, GUIDE would use the information to produce more specific splits of the form "$X \leq c$ or $X \in$ `S`", where `S` is a subset of missing-value codes. Illustrative examples of these two features are given in the GUIDE manual (Loh 2018).

Acknowledgements The authors thank Tao Shen, Yu-Shan Shih and Shijie Tang for their helpful comments. Data used in the preparation of this article were obtained from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. As such, the following organizations and individuals within the PRO-ACT Consortium contributed to the design and implementation of the PRO-ACT Database and/or provided data, but did not participate in the analysis of the data or the writing of this report: Neurological Clinical Research Institute, MGH Northeast ALS Consortium Novartis Prize4Life Israel Regeneron Pharmaceuticals, Inc., and Sanofi Teva Pharmaceutical Industries, Ltd.

References

Aalen O (1978) Nonparametric inference for a family of counting processes. Ann Stat 6:701–726
Ahn H, Loh W-Y (1994) Tree-structured proportional hazards regression modeling. Biometrics 50:471–485
Aitkin M, Clayton D (1980) The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM. Appl Stat 29:156–163
Atassi N, Berry J, Shui A, Zach N, Sherman A, Sinani E, Walker J, Katsovskiy I, Schoenfeld D, Cudkowicz M, Leitner M (2014) The PRO-ACT database: Design, initial analyses, and predictive features. Neurology 83(19):1719–1725
Breiman L, Friedman JH, Olshen RA, Stone CJ (1984) Classification and regression trees. Wadsworth, Belmont
Breslow N (1972) Contribution to the discussion of regression models and life tables by D. R. Cox. J R Stat Soc Ser B 34:216–217
Chan K-Y, Loh W-Y (2004) LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. J. Comput. Graph. Stat. 13:826–852
Chaudhuri P, Loh W-Y (2002) Nonparametric estimation of conditional quantiles using quantile regression trees. Bernoulli 8:561–576
Chaudhuri P, Huang M-C, Loh W-Y, Yao R (1994) Piecewise-polynomial regression trees. Stat Sin 4:143–167
Chaudhuri P, Lo W, Loh W, Yang C (1995) Generalized regression trees. Stat Sin 5:641–666
Dusseldorp E, Meulman JJ (2004) The regression trunk approach to discover treatment covariate interaction. Psychometrika 69:355–374
Foster JC, Taylor JMG, Ruberg SJ (2011) Subgroup identification from randomized clinical trial data. Stat Med 30:2867–2880
Gnanadesikan R (1997) Methods for statistical data analysis of multivariate observations, 2nd edn. Wiley, New York
Kim H, Loh W-Y (2001) Classification trees with unbiased multiway splits. J Am Stat Assoc 96:589–604
Kim H, Loh W-Y (2003) Classification trees with bivariate linear discriminant node models. J Comput Graph Stat 12:512–530
Laird N, Olivier D (1981) Covariance analysis of censored survival data using log-linear analysis techniques. J Am Stat Assoc 76:231–240
Lawless J (1982) Statistical models and methods for lifetime data. Wiley, New York
Lipkovich I, Dmitrienko A (2014) Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES. J Biol Stand 24:130–153
Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) Subgroup identification based on differential effect search — a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 30:2601–2621
Loh W-Y (1987) Calibrating confidence coefficients. J Am Stat Assoc 82:155–162
Loh W-Y (1991a) Bootstrap calibration for confidence interval construction and selection. Stat Sin 1:477–491
Loh W-Y (1991b) Survival modeling through recursive stratification. Comput Stat Data Anal 12:295–313
Loh W-Y (2002) Regression trees with unbiased variable selection and interaction detection. Stat Sin 12:361–386
Loh W-Y (2006) Regression tree models for designed experiments. In: Rojo J (ed) Second E. L. Lehmann Symposium, vol 49. IMS lecture notes-monograph series. Institute of Mathematical Statistics, pp 210–228
Loh W-Y (2009) Improving the precision of classification trees. Ann Appl Stat 3:1710–1737
Loh W-Y (2014) Fifty years of classification and regression trees (with discussion). Int Stat Rev 34:329–370
Loh W-Y (2018) GUIDE user manual. University of Wisconsin, Madisons

Loh W-Y, Shih Y-S (1997) Split selection methods for classification trees. Stat Sin 7:815–840
Loh W-Y, Vanichsetakul N (1988) Tree-structured classification via generalized discriminant analysis (with discussion). J Am Stat Assoc 83:715–728
Loh W-Y, Zheng W (2013) Regression trees for longitudinal and multiresponse data. Ann Appl Stat 7:495–522
Loh W-Y, He X, Man M (2015) A regression tree approach to identifying subgroups with differential treatment effects. Stat Med 34:1818–1833
Loh W-Y, Fu H, Man M, Champion V, Yu M (2016) Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. Stat Med 35:4837–4855
Loh W-Y, Cao L, Zhou P (2019a) Subgroup identification for precision medicine: a comparative review of thirteen methods. Data Min Knowl Disc **9**(5):e1326
Loh W-Y, Eltinge J, Cho MJ, Li Y (2019b) Classification and regression trees and forests for incomplete data from sample surveys. Stat Sin 29:431–453
Loh W-Y, Man M, Wang S (2019c) Subgroups from regression trees with adjustment for prognostic effects and post-selection inference. Stat Med 38:545–557
Morgan JN, Sonquist JA (1963) Problems in the analysis of survey data, and a proposal. J Am Stat Assoc 58:415–434
Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin J (2005) Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. Stat Comput 15:231–239
Seibold H, Zeileis A, Hothorn T (2016) Model-based recursive partitioning for subgroup analyses. Int J Biostat 12(1):45–63
Su X, Tsai C, Wang H, Nickerson D, Bogong L (2009) Subgroup analysis via recursive partitioning. J Mach Learn Res 10:141–158
Therneau T, Atkinson B (2018) rpart: recursive partitioning and regression trees. R package version 4.1-13
Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. J Comput Graph Stat 17:492–514