Subgroups from regression trees with adjustment for prognostic effects and post-selection inference*

Wei-Yin Loh
loh@stat.wisc.edu
University of Wisconsin
Madison, WI 53706, U.S.A.

Michael Man
man_michael@lilly.com
Eli Lilly and Company
Indianapolis, IN 46285, U.S.A.

Shuaicheng Wang
swa@BioStatSolutions.com
BioStat Solutions, Inc.
Frederick, MD 21703, U.S.A.

Abstract

Identification of subgroups with differential treatment effects in randomized trials is attracting much attention. Many methods employ regression tree algorithms. This article addresses two important questions arising from the subgroups. How to ensure that treatment effects in subgroups are not confounded with effects of prognostic variables? How to determine the statistical significance of treatment effects in the subgroups? We address the first question by selectively including linear prognostic effects in the subgroups in a regression tree model. The second question is more difficult because it falls within the subject of post-selection inference. We use a bootstrap technique to calibrate normal-theory t-intervals so that their expected coverage probability, averaged over all the subgroups in a fitted model, approximates the desired confidence level. It can also provide simultaneous confidence intervals for all subgroups. The first solution is implemented in the GUIDE algorithm and is applicable to data with missing covariate values, two or more treatment arms, and outcomes subject to right censoring. Bootstrap calibration is applicable to any subgroup identification method; it is not restricted to regression tree models. Two real examples are used for illustration: a diabetes trial where the outcomes are completely observed but some covariate values are missing, and a breast cancer trial where the outcome is right-censored.

Key words: bootstrap; differential treatment effect; missing value; precision medicine.

1 Introduction

Heterogeneity in patient populations often requires different therapies to be prescribed to different individuals. Precision medicine seeks to identify patient subgroups, defined by

*To appear in *Statistics in Medicine*

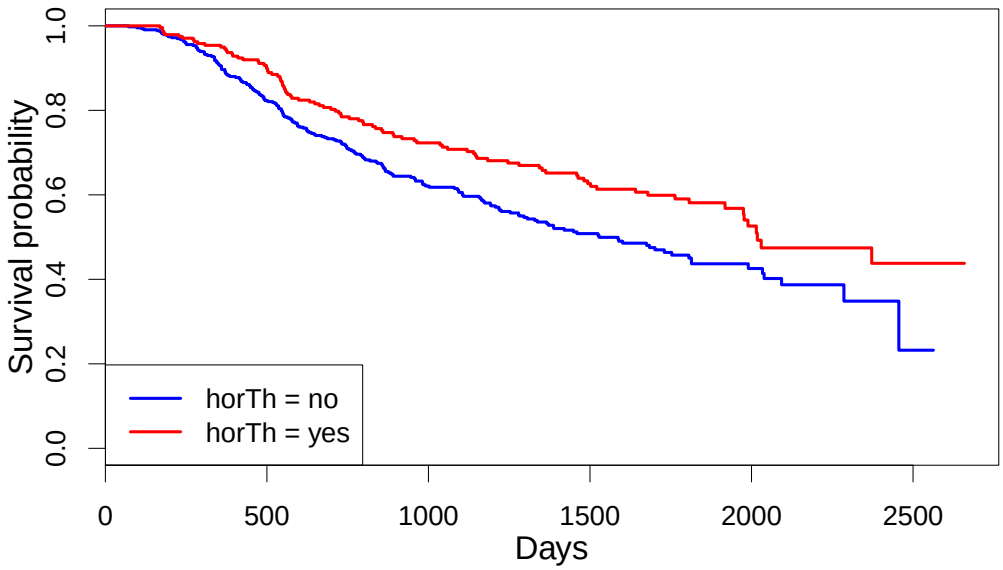

Figure 1: Kaplan-Meier survival curves for all subjects in breast cancer data

patient characteristics, that exhibit differential treatment effects. Regression tree methods can find subgroups by recursively partitioning the data [1–6]. Being primarily focused on detection of treatment interaction effects, the partitions typically employ predictive variables (variables that interact with the treatment) instead of prognostic variables (variables that predict the outcome regardless of treatment, such as age and disease stage at baseline) [7]. As a result, prognostic variables may be hidden and their effects confounded with treatment effects in the subgroups. For example, [8] warned that "efficacy assessments, as well as the assessment of benefit/risk, require careful inspection of relevant prognostic subgroups of clinical trials."

To illustrate, consider a German breast cancer study [9] where 686 subjects with primary node positive breast cancer were randomized to receive hormone therapy or not (`horTh`, yes, no). The response was recurrence-free survival time (8–2659 days; 299 uncensored, 387 censored) and there were 7 predictor variables: `age` (21–80 years), `tsize` (tumor size, 3–120 mm), `pnodes` (number of positive lymph nodes, 1–51), `progrec` (progesterone receptor status, 0–2380 fmol), `estrec` (estrogen receptor status, 0–1144 fmol), `menostat` (menopausal status, pre/post), and `tgrade` (tumor grade, 1, 2, 3). The Kaplan-Meier survival curves in Figure 1 show that, on average, hormone therapy increases recurrence-free survival time. Treatment remains statistically significant when adjusted for covariates in a proportional hazards model. The p-values in Table 1 reveal, however, that variables `tgrade`, `progrec` and, especially, `pnodes` are more significant and hence have strong prognostic effects.

An analysis using the GUIDE [10] regression tree method finds that subjects in the subgroup defined by "`progrec` ≤ 21" do not, on average, benefit from treatment [5]. Figure 2 shows the regression tree and the Kaplan-Meier curves in the subgroups (terminal nodes). Confidence intervals of relative risk for therapy versus no therapy, calculated using a bootstrap method described later, confirm that treatment is not statistically significant at the 0.05 level for `progrec` ≤ 21. As is quite typical, the result does not involve any prognostic variables.

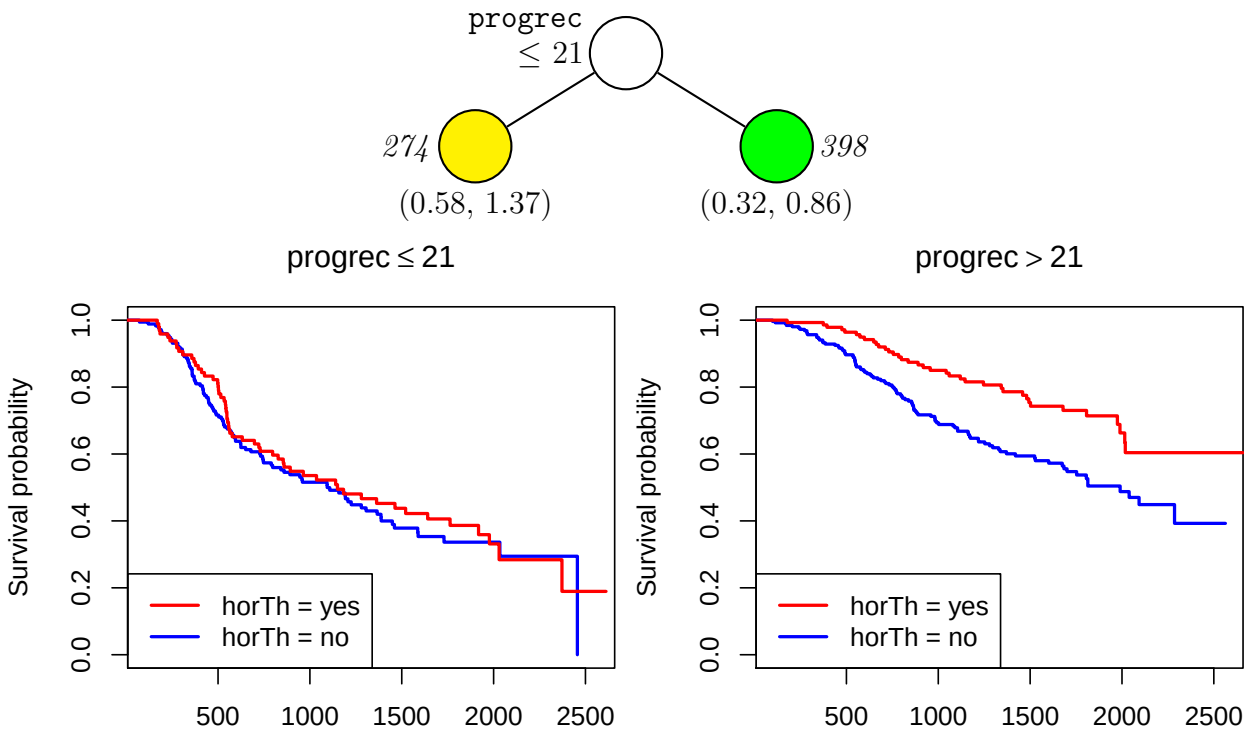

Figure 2: GUIDE regression tree for breast cancer data. At each intermediate node, a subject goes to the left subnode if and only if the condition is satisfied. Sample sizes are printed in italics beside nodes and 95% bootstrap confidence intervals of relative risks for therapy versus no therapy are printed below nodes.

Random treatment assignment ensures that the covariates are approximately balanced with respect to treatment at the root node of the tree. But the balance is not necessarily maintained at the terminal nodes. The possibility exists that the observed treatment effect in any terminal node is partly attributable to prognostic effects. A naïve way to account for prognostic effects is to regress them out with a fitted model and use the residuals to find subgroups. This is problematic in three respects. First, it depends on model choice. If the fitted model is linear in the covariates, then only linear prognostic effects may be removed. Besides, a variable may be prognostic in only one part of the sample space, e.g., for males but not females. Second, the residuals from the fitted model are not independent, which violates the assumption of independent observations most methods require. Third, residuals may not be readily defined when the outcomes are subject to censoring. One objective of this article is to solve these problems by extending GUIDE to explicitly include linear prognostic effects within subgroups.

Subgroup analysis, as carried out in the past, is known to have reproducibility issues, mainly due to multiple testing that inflate Type I error probability (probability of finding a subgroup where none exists). Many recommendations have been proposed to deal with them, including pre-specifying and limiting the number of subgroups, adjusting for multiple testing, and reporting all subgroups tested [11–16]. Given the power of machine learning algorithms to search for subgroups, it does not make sense to pre-specify them simply to enable multiplicity adjustment. Besides, the latter is difficult even for pre-specified subgroups because, as [17] noted, "it is inherently a multiple testing problem with the complication that test statistics for subgroups are typically highly dependent, making simple multiplicity corrections such as the Bonferroni correction too conservative." A second goal of this article is to extend a bootstrap method of multiplicity adjustment [6, 18, 19] that is ideally suited for computerized subgroup search: it does not require pre-specification of subgroups nor limits on the number of subgroups examined.

The remainder of this article is organized as follows. Section 2 briefly reviews some previous regression tree algorithms for subgroup identification. Section 3 presents the subgroup identification method, where treatment effects are adjusted for linear prognostic effects, for data with uncensored outcomes and applies it to data from a diabetes randomized trial. Section 4 describes the bootstrap calibration technique for interval estimation of treatment effects within subgroups. Section 5 extends the ideas to outcomes subject to censoring and applies them to the breast cancer data. Section 6 concludes the article with some remarks.

2 Previous regression tree methods

The first regression tree algorithm, called AID [20], appeared in 1963. It was followed two decades later by CART [21]. Both algorithms fit piecewise-constant regression models only, and search all splits on all variables to minimize the sum of squared residuals in the constant models fitted to the nodes induced by each split. This strategy gives preferential bias to variables that allow more splits [10, 22–25]. Furthermore, being piecewise-constant models, AID and CART are unsuitable for subgroup identification when there is a treatment variable. Either the variable is chosen to split a node, in which event the treated subjects go to one node and the untreated to the other, or the treatment variable is not selected at all. What

is needed is a method that fits a linear model (linear in treatment effects) in each node. The first algorithm to do this appears to be [26].

Let $\mathbf{X} = (X_1, X_2, \ldots, X_K)$ denote a K-dimensional vector of covariate values and let $(U_1, \mathbf{X}_1)$, $(U_2, \mathbf{X}_2)$, $\ldots$, $(U_n, \mathbf{X}_n)$ be the survival times and covariate vector values of n subjects. For subject i, let S_i be an independent observation from some censoring distribution and let $\delta_i = I(U_i < S_i)$ be the event indicator. The observed data vector of subject i is $(Y_i, \delta_i, \mathbf{X}_i)$, where $Y_i = \min(U_i, S_i)$. Let $\lambda(y, \mathbf{x})$ denote the hazard function at time y and covariate vector $\mathbf{x}$. The proportional hazards model postulates that $\lambda(y, \mathbf{x}) = \lambda_0(y) \exp(\eta)$, where $\lambda_0(y)$ is a baseline hazard function independent of $\mathbf{x}$, and $\eta = \beta' \mathbf{x}$ is a linear function of the covariates. Assuming that the treatment variable Z takes values 0 and 1, the method in [26] recursively partitions the data in a set t into subsets t_L and $t_R = t - t_L$. The set $t_L = \{X_j \le c_j\}$ for some constant c_j if X_j is an ordinal variable; otherwise $t_L = \{X_j \in A_j\}$ for some subset A_j of values of X_j if the latter is categorical (i.e., unordered). The value of c_j or A_j is chosen to maximize the Cox partial likelihood ratio statistic for testing the hypothesis $H_0 : \lambda(y, \mathbf{x}) = \lambda_{0t} \exp(\beta_0 z I(\mathbf{x} \in t)\}$ versus $H_1 : \lambda(y, \mathbf{x}) = \lambda_{0t}(y) \exp\{\beta_1 z I(\mathbf{x} \in t_L) + \exp\{\beta_2 z I(\mathbf{x} \in t_R)\}$. The *Interaction Trees* (IT) method [1, 27] chooses instead the split that minimizes the p-value for testing $H_0 : \beta_3 = 0$ in the model $\lambda(y, \mathbf{x}) = \lambda_{0t}(y) \exp\{\beta_1 z + \beta_2 I(\mathbf{x} \in t_L) + \beta_3 z I(\mathbf{x} \in t_L)\}$ fitted to the data in t (note that $\lambda_{0t}(y)$ is a function of t and y). The *Virtual twins* (VT) method [2] first uses a random forest [28] model to estimate the treatment effect for each subject. Then it fits a CART tree model to the estimated effects to find the subgroups. IT and VT are biased toward splitting on variables that allow more splits [22, 25]. The *SIDES* method [3, 29, 30] finds multiple alternative subgroups that yield the most improvement in a criterion such as p-value of treatment effect difference between t_L and t_R. For each split, the procedure is repeated on the child node with the larger treatment effect. SIDES uses heuristic and resampling-based adjustments are employed to control the probability of false discovery. It also uses multiplicity adjustments to reduce selection bias, but its effectiveness has not been studied. All three methods require prior imputation of missing covariate values (a difficult task by itself) and none extends easily to data with three or more treatment arms. For the breast cancer data, neither IT nor SIDES found any subgroup and VT is inapplicable to censored responses.

3 Uncensored outcomes

The basic GUIDE regression tree algorithm recursively partitions the data and sample space using one predictor variable at a time. Each node of the tree is split into two child nodes such that the total residual sum of squares (or, more generally, deviance), summed over the two child nodes, is minimum. A node is not partitioned if its sample size falls below a pre-specified small value or its residual deviance is zero. The resulting tree is pruned to a smaller size using a cross-validation method similar to that in CART. GUIDE differs from CART, however, in two important respects: split selection and node modeling. Instead of searching all splits on all variables, GUIDE uses significance tests to select the most significant split variable and then finds the optimal split based on that variable. Besides saving substantial computation, this approach yields unbiased variable selection. It is this computational advantage that makes fitting a nontrivial model in each node practical.

Let Y be an uncensored response variable and let the treatment variable Z take values $1, 2, \ldots, G$. For subgroup identification, GUIDE fits the treatment-only least-squares model

$$EY = \beta_{t0} + \sum_{z=2}^{G} \beta_{tz} I(Z = z) \tag{1}$$

to each node t of the tree [5]. Differences $(\beta_{t_L 0} - \beta_{t_R 0})$ between sibling nodes t_L and t_R reflect the prognostic effect of the variable selected to split their parent node. To avoid selection bias, GUIDE uses significance tests to find the best X_j to split each node t before looking for the best c_j or A_j for the selected X_j [10]. There are two options to find the best X_j [5]. One, called *Gs*, computes a chi-squared test of the data for each treatment level, with the signs of the residuals as rows and the values of X_j (or discretized values of X_j if the latter is an ordinal variable) as columns. Then it converts each chi-squared statistic into one with a single degree of freedom, sums the single-degree of freedom chi-squareds over the treatment levels, and picks the X_j with the largest sum. The second option, called *Gi*, tests for lack of fit of an additive model fitted to the data in the node. Specifically, given a candidate split variable X, let $V = X$ if it is categorical; otherwise if X is ordinal, let V be the categorical variable obtained by discretizing the X values at the sample quantiles into H groups, where $H = 3$ or 4, depending on whether the number of observations in the node is less or greater than $30G$. If X has missing values one group is reserved for missing values. The additive model $EY = \beta_{t0} + \sum_{z=2}^{G} \beta_{tz} I(Z = z) + \sum_{v=1}^{H} \gamma_{tv} I(V = v)$ is tested against the full model $EY = \eta_{t0} + \sum_{z=2}^{G} \sum_{v=1}^{H} \omega_{tvz} I(V = v, Z = z)$ and the X variable with the smallest lack-of-fit p-value selected to split the node. We focus on the Gi approach here because it is more sensitive to predictive variables.

Although model (1) accounts for some prognostic effects through the β_{t0} parameters, they are limited to the split variables, which (by design of Gi) tend to be predictive rather than prognostic. This is not a problem if every variable has both predictive and prognostic effects. But if there are prognostic variables that do not have predictive effects, the prognostic effects may be confounded with the treatment effects within nodes. In principle, it is straightforward to allow for the effects of prognostic variables by simply including them as linear predictors in each node. The question is how many and which, if not all, variables to use. Including all variables is potentially problematic because the sample sizes in the nodes shrink as they are partitioned. If there are many variables, this will limit the number of splits and thus reduce the probability of finding subgroups. Our solution is to fit only a single prognostic variable at each node. Specifically, model (1) is replaced by the linear prognostic model

$$EY = \beta_{t0} + \beta_{t1} X^{(t)} + \sum_{z=2}^{G} \beta_{tz} I(Z = z) \tag{2}$$

where $X^{(t)}$ is the prognostic variable minimizing the sum of squared residuals in node t. Two obvious advantages of this approach are execution speed and avoidance of computational difficulties due to multicollinearity in the X variables. A third is the possibility of graphing the fitted model and data in each subgroup (see below for an example). Although using only a single linear predictor may seem suboptimal, continued splitting followed by pruning of the nodes will likely result in the prognostic variables being correlated in the terminal nodes

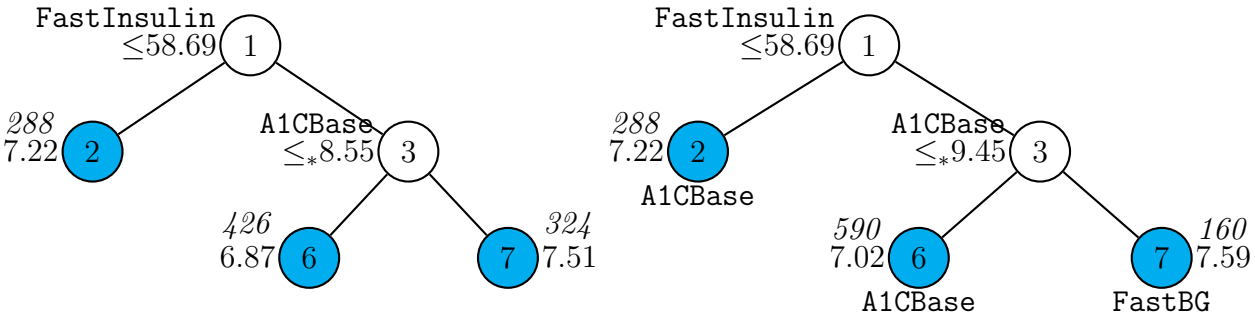

Figure 3: GUIDE models without (left) and with (right) linear prognostic control for diabetes data. At each split, an observation goes to the left branch if and only if the condition is satisfied. The symbol '$\leq_*$' stands for '$\leq$ or missing'. Beside each node are the sample size (in *italics*) and mean `A1C10`. The variable beneath each terminal node in the tree on the right is the best linear predictor.

so that inclusion of one with the strongest effect may be sufficient to account for the effects of the others.

We illustrate this idea on data from a multi-center, randomized double-blind trial on the long-term efficacy and safety of *Pioglitazone* vs *Gliclazide* in patients with Type 2 diabetes mellitus that is inadequately controlled by diet alone [31]. Gliclazide increases the amount of insulin produced by the pancreas while Pioglitazone is an "insulin sensitizer", i.e., it improves the ability of the body to use insulin. The trial consisted of 1249 subjects between 35 and 75 years old with HbA1c between 7.5% and 11.0% and for whom diet was prescribed for at least 3 months. Each subject was randomized to a 52-week treatment period consisting of a 16-week forced-titration period to a maximum dose and a 36-week maintenance period at the maximum tolerated dose of the drug. The treatments were 80mg Gliclazide (625 subjects), 30mg Pioglitazone (114 subjects), and 45mg Pioglitazone (510 subjects). Twenty-three baseline variables were measured for each subject; see Table 2 for their names and numbers of missing values. The response variable was HbA1c, measured for each subject at -2, 0, 4, 8, 12, 16, 24, 32, 42, and 52 weeks from baseline. For this illustration, we combine the 30mg and 45mg Pioglitazone groups into one "Pioglitazone" treatment group and take as the response variable `A1C10`, the HbA1c value at the tenth observation period (52 weeks). This yields a sample size of 1038 subjects; see [6] for an analysis of HbA1c over all time points.

The left side of Figure 3 shows the GUIDE result where a treatment-only model (1) is fitted to each node. It splits first on `FastInsulin`; if `FastInsulin` > 58.69, the tree splits further on `A1CBase`. The sample size and mean `A1C10` are printed beside each node. Neither IT nor SIDES finds any subgroups. The unpruned IT tree split first on `FastInsulin` ≤ 58.34, which is almost the same as that in the GUIDE tree. On the other hand, the initial subgroup found by SIDES before it failed to be confirmed was $\{$`BMI` ≤ 25.2, `Age` > 52, `FastBG` $\leq 12.8\}$.

The right side of Figure 3 shows the GUIDE model where a linear prognostic variable is fitted in each node. It is almost the same as the tree on the left side, the only difference being the split value at node 3. `A1CBase` is the best prognostic variable in nodes 2 and 6 and `FastBG` is the best in node 7. Figure 4 graphs the data and fitted model in the terminal

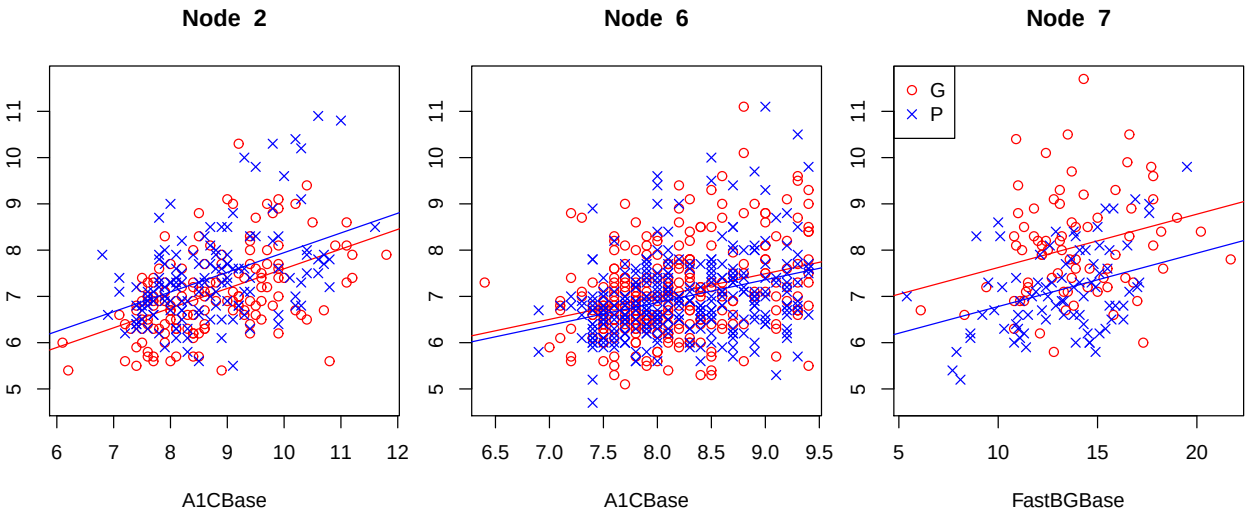

Figure 4: Plots of `A1C10` vs linear prognostic variable in nodes of tree on right side of Figure 3

nodes, with Gliclazide and Pioglitazone in red and blue colors, respectively. The slope of the parallel lines in each graph indicate the linear prognostic effect and the distance between the lines is the estimated treatment effect.

Table 3 shows the regression coefficients, t-statistics, and p-values in the linear models fitted to the nodes of the tree. The p-values and associated confidence intervals cannot be interpreted at face value, as they are computed assuming that the subgroups are pre-defined. Therefore the p-values are biased low and the intervals are too short. Correcting the bias using Bonferroni corrections is hard [17], because the subgroups change with each replication of the experiment. Instead, [6] used a bootstrap method to calibrate the nominal coverage of the confidence intervals. The basic idea of bootstrap calibration was originally conceived in [18, 19]. We review it and then extend it to the current problem in the next section.

4 Bootstrap calibration

Let F denote the population and let $\mathcal{D}$ denote the data which are assumed to be a random sample from F. Let $\mathcal{S}$ be the set of subgroups (terminal nodes) derived from $\mathcal{D}$. Given any $t \in \mathcal{S}$, let β_t denote the true treatment effect in t. Clearly, β_t is a function of F and $\mathcal{D}$. Given a desired value of $\alpha \in (0,1)$, we want to construct $(1-\alpha)$ confidence intervals, I_t say, such that "$P_F(\beta_t \in I_t) = 1-\alpha$ for each $t \in \mathcal{S}$." The quoted statement does not make sense of course, because $\mathcal{S}$ is not fixed. Nevertheless, suppose we proceed to construct a standard t-interval for the treatment effect in each node t. Let $T_{\nu,\alpha}$ denote the upper-α quantile of the t-distribution with ν degrees of freedom. Given any $\mathcal{D}'$, $\mathcal{S}'$ and $\alpha' \in (0,1)$, where $\mathcal{S}'$ is a set of subgroups constructed from $\mathcal{D}'$, define

$$J(\mathcal{D}', t', \alpha') = (\hat{\beta}_{t'} - T_{\nu_{t'},\alpha'/2}\,\hat{\sigma}_{t'},\ \hat{\beta}_{t'} + T_{\nu_{t'},\alpha'/2}\,\hat{\sigma}_{t'}), \quad t' \in \mathcal{S}' \tag{3}$$

where $\hat{\beta}_{t'}$ and $\hat{\sigma}_{t'}$ are the treatment effect estimate and standard error and $\nu_{t'}$ the residual degrees of freedom, computed from the $\mathcal{D}'$ observations in $t' \in \mathcal{S}'$. The interval $J(\mathcal{D}, t, \alpha)$

is just the usual $(1-\alpha)$ t-interval for the treatment effect in subgroup $t \in \mathcal{S}$ assuming t is pre-specified. It is too short (because $\hat{\beta}_t$ may not have a t-distribution and $\hat{\sigma}_t^2$ most likely underestimates the error variance) but the interval may still provide a good approximation if we can widen it by an appropriate amount, either by increasing $\hat{\sigma}_t^2$ or decreasing the value of α. The first solution was proposed in [5], which replaces $\hat{\sigma}_t^2$ with a bootstrap estimate. Although simulation results reported there show that the resulting intervals have better coverage probabilities, it is not clear that the bootstrap variance estimates are consistent. We now propose an alternative solution that keeps the variance estimates unchanged but widens the intervals by decreasing the value of α. Besides being more intuitive, its applicability is more general and is not limited to t-intervals.

Consider first the hypothetical situation where we know F and can repeatedly draw random samples from it. Let $\mathcal{D}_i$ denote the ith sample ($i = 1, 2, \ldots, L$, for some large integer L). Using the same algorithm that produced $\mathcal{S}$ from $\mathcal{D}$, obtain a new set of subgroups $\mathcal{S}_i$ from $\mathcal{D}_i$. Given α', apply (3) to the $\mathcal{D}_i$ observations to construct intervals $J(\mathcal{D}_i, t^*, \alpha')$ for $t^* \in \mathcal{S}_i$. Let $\beta_{t^*,F}$ denote the true treatment effect in subgroup $t^* \in \mathcal{S}_i$ *computed from* F. Let $|\mathcal{S}_i|$ be the number of subgroups in $\mathcal{S}_i$ and m_i be the number of intervals for which $\beta_{t^*,F} \in J(\mathcal{D}_i, t^*, \alpha')$. Define $\gamma_i(\alpha') = m_i/|\mathcal{S}_i|$ and $\bar{\gamma}(\alpha') = L^{-1}\sum_{i=1}^{L}\gamma_i(\alpha')$. Repeat the whole exercise with different values of α' to find α_F such that $\bar{\gamma}(\alpha_F) = 1-\alpha$. Now use (3) with $\mathcal{D}' = \mathcal{D}$, $\mathcal{S}' = \mathcal{S}$ and $\alpha' = \alpha_F$ to construct the *calibrated* intervals $J(\mathcal{D}, t, \alpha_F)$ for $t \in \mathcal{S}$. Then

$$E_F\Big\{|\mathcal{S}|^{-1}\sum_{t\in\mathcal{S}} I(\beta_{t,F} \in J(\mathcal{D}, t, \alpha_F))\Big\} = 1-\alpha \tag{4}$$

where the expectation is over all random samples $\mathcal{D}$ from F. That is, the coverage probability of $J(\mathcal{D}, t, \alpha_F)$, averaged over the subgroups in the tree model constructed from $\mathcal{D}$, has expected value $(1-\alpha)$. If F is unknown, we replace it with $\hat{F}$, the empirical distribution of $\mathcal{D}$, in the procedure to obtain the *bootstrap-calibrated* intervals $J(\mathcal{D}, t, \alpha_{\hat{F}})$, where $\bar{\gamma}(\alpha_{\hat{F}}) = 1-\alpha$.

Algorithm 1 gives the steps formally, including those for bootstrap calibration of *simultaneous* confidence intervals (with θ referring to simultaneous coverage). The values of $\alpha_{\hat{F}}$ for the non-simultaneous and simultaneous intervals are linearly interpolated between the pair of α' values whose bootstrap coverage probabilities are just above and below the desired levels. Figure 5 plots the coverage values, based on 1000 bootstrap iterations, of the non-simultaneous and simultaneous intervals for the diabetes data for a grid of nominal α values. The values of $\alpha_{\hat{F}}$ for 95% non-simultaneous and 90% simultaneous intervals are 0.00393 and 0.00035, respectively. Therefore the unadjusted p-values in Table 3 are statistically significant at the 0.05 non-simultaneous and 0.10 simultaneous levels only if they are less than 0.00393 and 0.00035, respectively. Figure 6 adds bootstrap-calibrated 95% non-simultaneous and 90% simultaneous confidence intervals for the treatment effects to the trees in Figure 3. Based on 95% non-simultaneous intervals, treatment is statistically significant in nodes 2 and 7 with and without adjusting for prognostic effects. But based on 90% simultaneous intervals, treatment is significant in node 7 only after prognostic adjustment.

Data: Given $K > 0$ and $\alpha \in (0,1)$, $\alpha_1 < \alpha_2 < \ldots < \alpha_K = \alpha$; $Z_i \in \{1, 2, \ldots, G\}$; tree with nodes $t_1, t_2, \ldots, t_L$ constructed from $\mathcal{D} = \{(\mathbf{X}_i, \mathbf{Y}_i, Z_i), i = 1, 2, \ldots, n\}$.

Result: $(1-\alpha)$ individual and simultaneous confidence intervals for β_{tz} for $t = t_1, t_2, \ldots, t_L$ and $z = 2, \ldots, G$.

begin

$\gamma_k \leftarrow 0$ for $k = 1, 2, \ldots, K$; `/* individual coverage probabilities */`

$\theta_k \leftarrow 0$ for $k = 1, 2, \ldots, K$; `/* simultaneous coverage probabilities */`

for $b \leftarrow 1$ **to** B **do**

bootstrap $\mathcal{D}_b^* = \{(\mathbf{X}_i^*, \mathbf{Y}_i^*, Z_i^*), i = 1, 2, \ldots, n\}$ from $\mathcal{D}$;

construct tree from $\mathcal{D}_b^*$ with nodes $t_{b1}^*, t_{b2}^*, \ldots, t_{bL_b}^*$;

for $z \leftarrow 2$ **to** G **do**

for $l \leftarrow 1$ **to** L_b **do**

$M^* \leftarrow$ model (2) fitted to $\mathcal{D}_b^*$ cases belonging to partition t_{bl}^*;

$\hat{\beta}(t_{bl}^*, z) \leftarrow$ value of β_{tz} from M^*;

$M \leftarrow$ model (2) fitted to $\mathcal{D}$ cases belonging to partition t_{bl}^*;

$\beta(t_{bl}^*, z) \leftarrow$ value of β_{tz} from M (true model);

for $k \leftarrow 1$ **to** K **do**

$I_{klz} \leftarrow (1-\alpha_k)$ nominal interval $\hat{\beta}(t_{bl}^*, z) \pm z_{\alpha_k/2} SE(\hat{\beta}(t_{bl}^*, z))$;

if $\beta(t_{bl}^*, z) \in I_{klz}$ **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**

$\gamma_k \leftarrow \gamma_k + \{(G-1)L_b\}^{-1} \sum_l \sum_z c_{klz}$;

if $\min_{lz} c_{klz} = 1$ **then**

$\theta_k \leftarrow \theta_k + 1$

end

end

end

$\gamma_k \leftarrow \gamma_k / B$ and $\theta_k \leftarrow \theta_k / B$ for $k = 1, 2, \ldots, K$;

$p \leftarrow$ smallest k such that $\gamma_k < 1 - \alpha$; $q \leftarrow$ smallest k such that $\theta_k < 1 - \alpha$;

$f \leftarrow (\gamma_{p-1} - 1 + \alpha)/(\gamma_{p-1} - \gamma_p)$; $g \leftarrow (\gamma_{q-1} - 1 + \alpha)/(\gamma_{q-1} - \gamma_q)$;

$\alpha' \leftarrow (1-f)\alpha_{p-1} + f\alpha_p$; $\alpha'' \leftarrow (1-g)\alpha_{q-1} + g\alpha_q$;

construct $(1-\alpha')$ non-simultaneous and $(1-\alpha'')$ simultaneous intervals for β_{tz} for $t = t_1, t_2, \ldots, t_L$; $z = 2, \ldots, G$;

end

Algorithm 1: Bootstrap calibration of confidence intervals for treatment effects

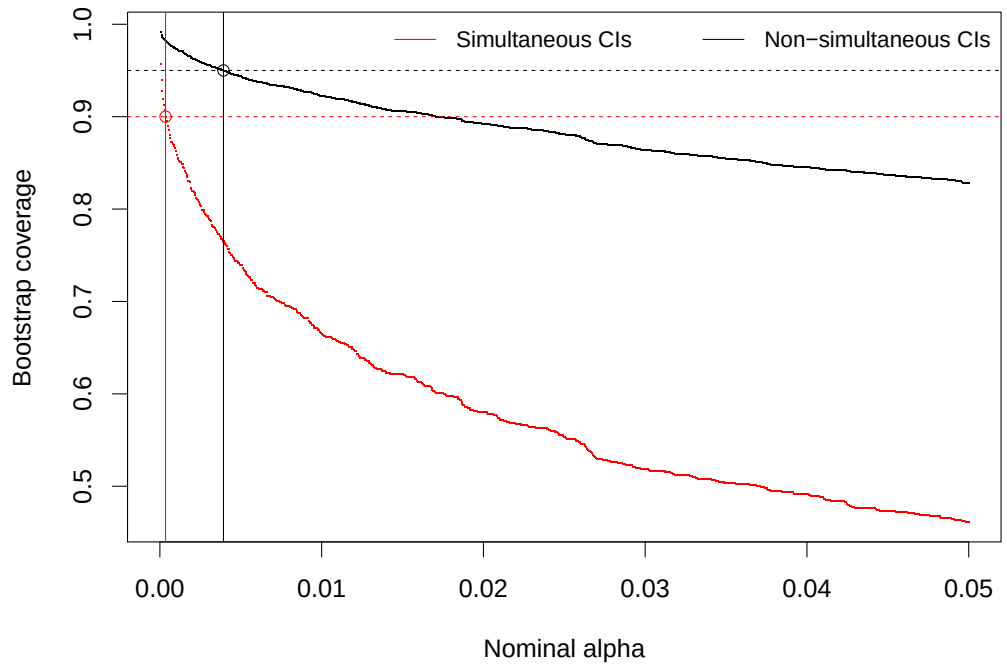

Figure 5: Coverage of simultaneous (red) and non-simultaneous (black) intervals of treatment effect with linear prognostic control for diabetes data, based on 1000 bootstrap iterations

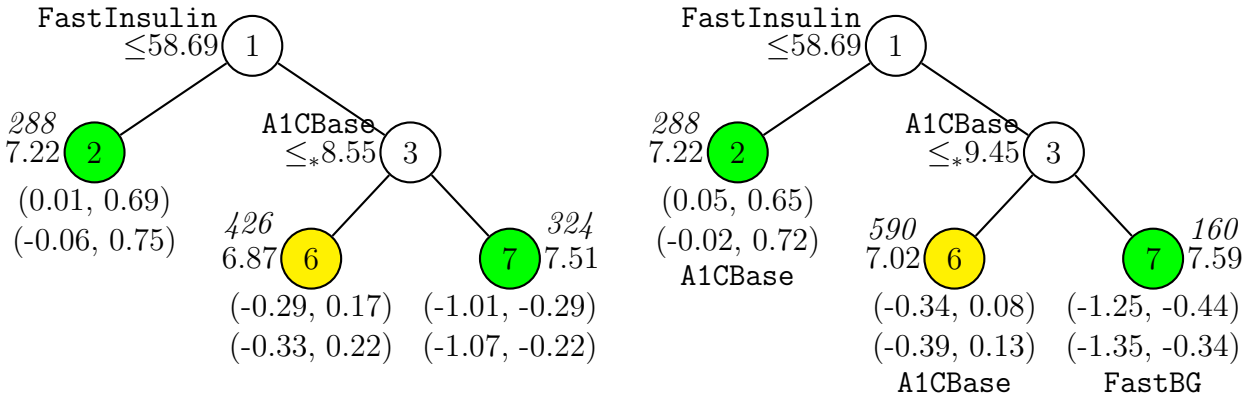

Figure 6: GUIDE models without (left) and with (right) linear prognostic control for diabetes data. Beside each node are the sample size (in *italics*) and mean `A1C10`. Below each node are 95% non-simultaneous (top) and 90% simultaneous (bottom) bootstrap-calibrated intervals for treatment effect (Pioglitazone – Gliclazide). The best linear prognostic variable is printed below each terminal node in the tree on the right. Nodes with statistically significant and non-significant treatment effects at the 95% non-simultaneous level are colored green and yellow, respectively.

5 Censored outcomes

If the response variable is subject to right censoring, we fit the proportional hazards model

$$\log \lambda(y, \mathbf{x}) = \log \lambda_0(y) + \beta_{t0} + \sum_{z=2}^{G} \beta_{tz} I(Z = z) \tag{5}$$

to each node t instead of the linear model (1). The usual proportional hazards model does not have a constant term because it can be absorbed in $\lambda_0(y)$. Here the β_{t0} term is needed to represent the effect of node t. As a result, $\lambda_0(y)$ is defined only up to a multiplicative constant and the β_{t0} values are over-parameterized, although their contrasts are well defined. Given a split by a variable X of a node into child nodes t_L and t_R, the difference $\beta_{t_L 0} - \beta_{t_R 0}$ is a measure of the prognostic effect of X.

Unlike IT and other survival tree methods [1, 26], GUIDE uses the same baseline hazard function $\lambda_0(y)$ for all t to ensure that the model as a whole has proportional hazards:

$$\log \lambda(y, \mathbf{x}) = \log \lambda_0(y) + \sum_{t} \Big\{ \beta_{t0} + \sum_{z=2}^{G} \beta_{tz} I(Z = z) \Big\}. \tag{6}$$

This is carried out by using Poisson regression to estimate the regression coefficients in model (6) via a well-known connection between the proportional hazard likelihood and the Poisson likelihood [5, 32, 33]. Specifically, let $\Lambda_0(y) = \int_{-\infty}^{y} \lambda_0(u)\, du$ denote the baseline cumulative hazard function. The regression coefficients of the proportional hazards model are obtained by iteratively fitting a Poisson regression tree [34] to the data, using the event indicators δ_i as Poisson responses and $\log \Lambda_0(y_i)$ as offset variable. At the first iteration, $\Lambda_0(y_i)$ is estimated by the Nelson-Aalen method [35, 36]. Thereafter, 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., [37, p. 361]).

Therefore in place of model (2), the data in each node are fitted with the Poisson model

$$\log E(\delta) = \log \Lambda_0(y) + \beta_{t0} + \beta_{t1} X^{(t)} + \sum_{z=2}^{G} \beta_{tz} I(Z = z). \tag{7}$$

Selection of a variable X to split each node follows the procedure in Section 3 with least squares replaced by Poisson regression. In the Gi method, for example, the variable X selected to split a node t is the one with the smallest lack-of-fit p-value in testing the Poisson model $\log E(\delta) = \log \Lambda_0(y) + \beta_{t0} + \sum_{z=2}^{G} \beta_{tz} I(Z = z) + \sum_v \gamma_{tv} I(V = v)$ against the full model $\log E(\delta) = \log \Lambda_0(y) + \eta_{t0} + \sum_{z=2}^{G} \sum_v \omega_{tvz} I(V = v, Z = z)$, where V is a categorical version of X.

The breast cancer example gives a trivial tree with no splits after pruning when a linear prognostic variable is included in each node. The best variable to split the root node, if it were split, is `estrec`. To induce a split, we re-fit the data with variable `estrec` excluded. Now a nontrivial tree is obtained with the 0-SE pruning rule (the 0-SE tree is the one with the smallest cross-validation estimate of deviance). The result is given in Figure 7 which also shows the tree model without controlling for linear prognostic effects. The tree structures

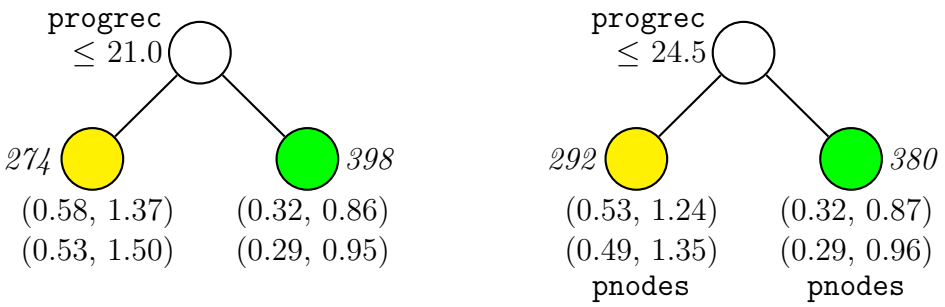

Figure 7: GUIDE proportional hazards regression trees for breast cancer data without (left) and with (right) linear prognostic control. At each split, an observation goes to the left branch if and only if the condition is satisfied. Beside each node is the sample size (in *italics*). Below each node are 95% non-simultaneous (top) and 90% simultaneous (bottom) bootstrap-calibrated intervals of relative risk due to treatment and the best linear prognostic variable. Nodes with statistically significant and non-significant treatment effects at the 95% non-simultaneous level are painted green and yellow, respectively.

are almost identical, except a slight difference in the split points. Variable `pnodes` is the best linear prognostic variable in both nodes of the tree on the right. Figure 8 plots the coverage probabilities of the unadjusted simultaneous and non-simultaneous intervals (based on 500 bootstraps) for a grid of 1000 α values. The bootstrap-calibrated value $\alpha_{\hat{F}}$ for 95% non-simultaneous intervals is 0.00989 and that for 90% simultaneous intervals is 0.00207. The respective bootstrap-calibrated intervals of relative risk due to treatment are given beneath each node of the tree in Figure 7. Treatment remains significant after allowing for the prognostic effect of `pnodes` in the subgroup where `progrec` > 24.5.

A simulation experiment was carried out to check the coverage probability of the bootstrap-calibrated intervals for the breast cancer data. To maximize the relevance of the simulation to the real data, the latter were used as the true (discrete) population in the simulation, which was carried out as follows.

1. Draw a simple random sample with replacement from the real data.
2. Construct a GUIDE tree model with linear prognostic control on the sampled data.
3. Use Algorithm 1 to construct bootstrap-calibrated 90% simultaneous and 95% non-simultaneous intervals for the treatment effect in each subgroup of the model with 100 bootstrap iterations and an α-grid of 1000 equally-spaced values over the interval $(1/20000, 1/20)$.
4. Use the real (population) data to estimate the true treatment effect in each subgroup.
5. Find the proportion of intervals that contain the true values.
6. Repeat the above steps 1000 times to obtain the simulated coverage probabilities of the bootstrap intervals.

Table 5 shows the results together with those of the uncalibrated intervals and the simulation standard errors. The average coverage of the uncalibrated 95% non-simultaneous intervals

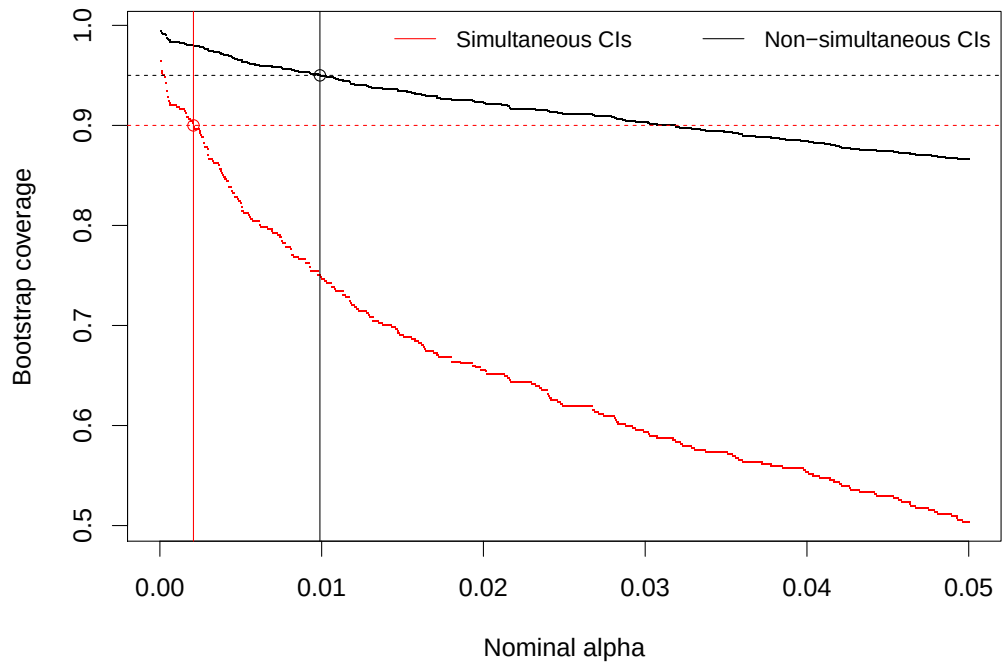

Figure 8: Coverage of simultaneous (red) and non-simultaneous (black) intervals of treatment effect with linear prognostic control for breast cancer data, based on 500 bootstrap iterations

is about 0.875 while that of the calibrated intervals is about 0.939 (the latter is within two simulation standard errors of 0.950). Similarly, the coverage of the calibrated 90% simultaneous intervals is 0.890, which is one simulation standard error from the target of 0.90. The average values of the bootstrap-calibrated $\alpha_{\hat{F}}$ over the 1000 simulation trials are 0.0187 and 0.0078 for the non-simultaneous and simultaneous intervals, respectively.

6 Conclusion

We pursued two main ideas in this article. The first is a flexible and nonparametric way to explicitly allow for the effects of prognostic variables in the search and identification of subgroups. Regressing out the linear effects of prognostic variables before searching for subgroups is undesirable for many reasons, the most important being: (1) the residuals after removing linear prognostic effects are not independent, which violates a common assumption of subgroup search algorithms; (2) it may not be easy to identify the prognostic variables, because a variable may be both prognostic and predictive; and (3) the effect of a prognostic variable may not be linear or uniform throughout the sample space. The last point is demonstrated by the diabetes example, where different variables are prognostic in different parts of the sample space.

Although we employed only one linear prognostic predictor in each node here, it is theoretically permissible (and possibly preferable) to allow more than one or all potentially prognostic variables. Using all prognostic variables (i.e., fitting a multiple linear model in each node) necessarily makes the tree shorter, because each node must contain correspondingly more observations. This problem can be avoided by using stepwise regression or a regu-

larization method such as LASSO [38], at the cost of greater computation time. Whether the additional expense yields increased precision in subgroup selection and treatment effect estimation is left for future research.

The second main idea here is post-selection inference. Until now this is performed in a post-hoc fashion, with the focus being control of Type I error probability, either by pre-specification of subgroups or intricate multiplicity adjustments. In theory, subgroup pre-specification should not be necessary; the investigator should not be prevented from analyzing subgroups revealed by the data. Pre-specification is only necessary to enable multiplicity correction.

Bootstrap calibration is an alternative to multiplicity adjustment. Unlike Bonferroni-type corrections that often require careful mathematical analysis, calibration is fully automatic. All that is needed is repeated application of the search algorithm on bootstrap samples of the data. It is crucial that the algorithm (but not the subgroups) remain the same for all bootstrap samples. This does not preclude human expert-guided ad hoc search, but it is unrealistic to expect an expert to repeatedly and independently analyze large numbers of bootstrap data sets. Therefore calibration is best done with a computer search algorithm ([39] used "principled" and "disciplined" to describe algorithmic search).

Asymptotic validity of bootstrap calibration for a regression tree model depends on three assumptions: (i) the empirical distribution $\hat{F}$ converges to the true distribution F, (ii) the coverage probability of the confidence intervals is a smooth function of F, and (iii) the partitions of the tree converge, as the sample size increases. The first assumption, convergence of $\hat{F}$ to F, is necessary for any reasonable method to work. While it may be possible to construct counterexamples where the second ("smoothness") assumption is violated, we expect that is satisfied in real applications. The third assumption can be satisfied by setting a minimum threshold on the proportion of training samples in each terminal node, to prevent the tree structure from growing without bound [34, 40].

Permutation tests have been used to control Type I error probability in subgroup search. They are fundamentally different and less versatile than bootstrap calibration. In the former, synthetic data sets are generated from the real data by randomly permuting the treatment labels to simulate the null hypothesis of no treatment effect. The search algorithm is applied to each synthetic data set and the fraction of them that yield subgroups is obtained. If the fraction exceeds a given α level, the real data are deemed to contain no subgroups. Although this approach controls the Type I error probability, it has low power if the treatment effect is positive in one subgroup and negative in another. A variant that permutes treatment-outcome pairs instead of only treatment labels essentially tests the null hypothesis of neither treatment nor prognostic effects [3]. Permutation tests are particularly hard to conceive for assessing significance of treatment effects after adjustment for linear prognostic effects in subgroups. Our bootstrap calibration approach goes beyond controlling Type I error probability—it gives confidence intervals for treatments effects in the subgroups.

The linear prognostic adjustment method described here is implemented in the GUIDE software, which may be obtained from `www.stat.wisc.edu/~loh/guide.html`.

7 Acknowledgments

The authors are grateful to Lei Shen for bringing the problem of linear prognostic control to their attention and to Haoda Fu for sharing the diabetes data with them. They also wish to thank two reviewers and an associate editor for comments that led to improvements in the manuscript. WY Loh was partially supported by NSF grant DMS-1305725, NIH grant 1P01CA180945-01, and a grant from Eli Lilly and Company. The research was partially completed while Loh was visiting the Institute for Mathematical Sciences, National University of Singapore in July 2017.

References

1. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data *Int J Biostat.* 2008;4.

2. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data *Stat Med.* 2011;30:2867–2880.

3. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search — A recursive partitioning method for establishing response to treatment in patient subpopulations *Stat Med.* 2011;30:2601–2621.

4. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions *Stat Med.* 2014;33:219-237.

5. Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects *Stat Med.* 2015;34:1818-1833.

6. Loh WY, Fu H, Man M, Champion V, Yu M. Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables *Stat Med.* 2016;35:4837-4855.

7. Italiano A. Prognostic or predictive? It's time to get back to definitions! *J Clin Oncol.* 2011;29:4718.

8. Koch GG, Schwartz TA. An Overview of Statistical Planning to Address Subgroups in Confirmatory Clinical Trials *J Biopharm Stat.* 2014;24:72 - 93.

9. Schumacher M, Baster G, Bojar H, et al. Randomized 2 $\times$ 2 trial evaluating hormonal treatment and the duration of chemotherapy in node-positive breast cancer patients *J Clin Oncol.* 1994;12:2086–2093.

10. Loh WY. Regression trees with unbiased variable selection and interaction detection *Stat Sin.* 2002;12:361-386.

11. Oxman AD, Guyatt GH. A consumer's guide to subgroup analyses *Ann Intern Med.* 1992;116:78-84.

12. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine—reporting of subgroup analyses in clinical trials *N Engl J Med.* 2007;357:2189-2194.

13. Dijkman B, Kooistra B, Bhandari M. How to work with a subgroup analysis *Can J Surg.* 2009;52:515-522.

14. Sun X, Briel M, Busse JW, et al. Credibility of claims of subgroup effects in randomised controlled trials: systematic review *BMJ.* 2012;344.

15. Tanniou J, Tweel I, Teerenstra S, Roes KCB. Subgroup analyses in confirmatory clinical trials: time to be specific about their purposes *BMC Med Res Methodol.* 2016;16:1-15.

16. Pletcher MJ, McCulloch CE. The challenges of generating evidence to support precision medicine *JAMA Intern Med.* 2017;177:561-562.

17. Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification *J Biopharm Stat.* 2014;24:110 - 129.

18. Loh WY. Calibrating confidence coefficients *J Am Stat Assoc.* 1987;82:155–162.

19. Loh WY. Bootstrap calibration for confidence interval construction and selection *Stat Sin.* 1991;1:477–491.

20. Morgan JN, Sonquist JA. Problems in the analysis of survey data, and a proposal *J Am Stat Assoc.* 1963;58:415–434.

21. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees.* Belmont, California: Wadsworth 1984.

22. Loh WY, Shih YS. Split selection methods for classification trees *Stat Sin.* 1997;7:815–840.

23. Kim H, Loh WY. Classification trees with unbiased multiway splits *J Am Stat Assoc.* 2001;96:589–604.

24. Kim H., Loh W.-Y.. Classification trees with bivariate linear discriminant node models *J Comput Graph Stat.* 2003;12:512-530.

25. Loh WY. Fifty years of classification and regression trees (with discussion) *Int Stat Rev.* 2014;34:329-370.

26. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JR. Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria *Stat Comput.* 2005;15:231–239.

27. Su X, Tsai CL, Wang H, Nickerson DM, Bogong L. Subgroup analysis via recursive partitioning *J Mach Learn Res.* 2009;10:141–158.

28. Breiman L. Random forests *Mach Learn.* 2001;45:5–32.

29. Lipkovich I, Dmitrienko A. Strategies for identifying predictive biomarkers and subgroups with enhanced treatment effect in clinical trials using SIDES *J Biopharm Stat.* 2014;24:130 - 153.

30. Riviere MK. *SIDES: Subgroup Identification Based on Differential Effect Search* 2017. R package version 1.11.

31. Charbonnel BH, Matthews DR, Schernthaner G, Hanefeld M, Brunetti P. A long-term comparison of Pioglitazone and Gliclazide in patients with Type 2 diabetes mellitus: a randomized, double-blind, parallel-group comparison trial *Diabet Med.* 2004;22:399-405.

32. Aitkin M, Clayton D. The fitting of exponential, Weibull and extreme value distributions to complex censored survival data using GLIM *Appl Stat.* 1980;29:156–163.

33. Laird N, Olivier D. Covariance analysis of censored survival data using log-linear analysis techniques *J Am Stat Assoc.* 1981;76:231–240.

34. Chaudhuri P, Lo WD, Loh WY, Yang CC. Generalized regression trees *Stat Sin.* 1995;5:641–666.

35. Aalen OO. Nonparametric inference for a family of counting processes *Ann Stat.* 1978;6:701–726.

36. Breslow N. Contribution to the discussion of regression models and life tables by D. R. Cox *J R Stat Soc B.* 1972;34:216–217.

37. Lawless JF. *Statistical Models and Methods for Lifetime Data.* New York: Wiley 1982.

38. Tibshirani R, Saunders M, Rosset S, Zhu J, Knight K. Sparsity and smoothness via the fused lasso *J R Stat Soc B.* 2005;67:91-108.

39. Dmitrienko A, Millen B, Lipkovich I. Multiplicity considerations in subgroup analysis *Stat Med.* 2017;36:4446-4454.

40. Chaudhuri P, Huang MC, Loh WY, Yao R. Piecewise-polynomial regression trees *Stat Sin.* 1994;4:143–167.

Table 1: Proportional hazards model fitted to breast cancer data

Variable	Coef	p-value	Variable	Coef	p-value
horTh=yes	-0.3372	8.9e-03	tsize	0.0078	0.0507
age	-0.0094	0.3111	pnodes	0.0499	1.7e-11
meno=Pre	-0.2673	0.1449	progrec	-0.0022	0.0001
tgrade	0.2803	0.0082	estrec	0.0002	0.7084

Table 2: Baseline predictor variables and their numbers of missing values for 1038 subjects in diabetes data with HbA1c at 10 weeks. HOMA stands for Homeostasis Model Assessment; B refers to beta cell function, IR to insulin resistance, and S to insulin sensitivity.

Variable	#Miss	Variable	#Miss
HDL	29	Age	0
LDL	129	Weight	0
Total cholesterol	28	BMI	0
Triglycerides	28	Waist	2
Creatinine	1	A1CBase (baseline HbA1c)	0
FastInsulin (fasting insulin)	114	HOMA-S	136
ALT (alanine aminotransferase)	2	HOMA-IR	136
AST (aspartate aminotransferase)	2	HOMA-B	136
GGT (γ-glutamyl transpeptidase)	1	Diastolic blood pressure	0
C-peptide	817	Systolic blood pressure	0
Diabetes duration	0	Pulse	0
FastBG (fasting blood glucose)	0		

Table 3: Regression estimates in terminal nodes of tree on right side of Figure 3. "Treatment.P" refers to the effect of Pioglitazone relative to Gliclazide. A negative value indicates that Pioglitazone reduces HbA1c more than Gliclazide. The p-values in the last column do not take the search algorithm into account. The bootstrap-calibrated threshold for 0.05-level significance is 0.00275; i.e., a treatment effect is statistically significant at the 0.05 level if its unadjusted p-value is less than 0.00275.

Node	Regressor	Coefficient	t-stat	Unadjusted p-value
2	A1CBase	0.4254	8.90	0.0000
	Treatment.P	0.3488	3.48	0.0006
6	A1CBase	0.4933	8.23	0.0000
	Treatment.P	-0.1298	-1.82	0.0692
7	FastBG	0.1154	3.92	0.0001
	Treatment.P	-0.8426	-5.27	0.0000

Table 4: Regression estimates in terminal nodes of tree on right side of Figure 7. In the table, "`horTh.yes`" refers to the effect of hormone therapy versus no hormone therapy. A negative coefficient implies that hormone therapy reduces the hazard rate. The p-values in the last column do not take the search algorithm into account. The bootstrap-calibrated threshold for 0.05-level non-simultaneous significance is 0.00989 and that for 0.10-level simultaneous significance is 0.00207.

Subgroup	Regressor	Coefficient	t-stat	Unadjusted p-value
`progrec` ≤ 24.5	`pnodes`	0.0868	8.35	0.0000
	`horTh.yes`	-0.2092	-1.27	0.2063
`progrec` > 24.5	`pnodes`	0.0399	3.61	0.0003
	`horTh.yes`	-0.6433	-3.30	0.0011

Table 5: Simulated coverage probabilities (SEs in parentheses) of 95% non-simultaneous and 90% simultaneous intervals for treatment effect with linear prognostic control, using breast cancer data as simulation population. Results based on 1000 simulation trials, 100 bootstraps per trial, and a calibration grid of 1000 equally-spaced α-values over the interval $(1/20000, 1/20)$.

	95% non-simultaneous	90% simultaneous
Uncalibrated t interval	0.875 (0.010)	—
Bootstrap calibrated interval	0.939 (0.008)	0.890 (0.010)