Published in final edited form as:

J Biopharm Stat. 2019 ; 29(6): 1082–1102. doi:10.1080/10543406.2019.1584204.

Look Before You Leap: Systematic Evaluation of Tree-based Statistical Methods in Subgroup Identification

Yang Liu^a, Xiwen Ma^b, Donghui Zhang^c, Lijiang Geng^a, Xiaojing Wang^a, Wei Zheng^d, Ming-Hui Chen^{a,*}

^aDepartment of Statistics, University of Connecticut, Storrs, Connecticut, U.S.A.

^bKaryopharm Therapeutics Inc., Newton, MA, U.S.A.

^cGlobal Biostatistics and Programming, Sanofi US, Bridgewater, New Jersey, U.S.A.

^dKehang (Suzhou) information and technology Ltd. Co., Suzhou, Jiangsu, China

Abstract

Subgroup analysis, as the key component of personalized medicine development, has attracted a lot of interest in recent years. While a number of exploratory subgroup searching approaches have been proposed, informative evaluation criteria and scenario-based systematic comparison of these methods are still underdeveloped topics. In this article, we propose two evaluation criteria in connection with traditional type I error and power concepts, and another criterion to directly assess recovery performance of the underlying treatment effect structure. Extensive simulation studies are carried out to investigate empirical performance of a variety of tree-based exploratory subgroup methods under the proposed criteria. A real data application is also included to illustrate the necessity and importance of method evaluation.

Keywords

GUIDE; Interaction tree; Qualitative interaction trees; T-AIC/T-BIC; Virtual twins

1 Introduction

The search of homogeneity and heterogeneity has been a constant topic in numerous scientific fields. The fundamental goal of precision medicine is to assign the right treatments to the right patients, in other words, it is often desirable to identify the patient group who would respond to an experimental treatment, either in a beneficial manner or the other, i.e., for whom the control is preferable to the experimental treatment in clinical studies. While investigations of the latter group play a significant role in drug safety studies, searching of a potential beneficial patient subgroup is usually of more interest in practice. In addition, as discussed in [Lipkovich et al., 2017], the treatment effects could be still enhanced for some patients even when the experimental treatment has positive effects for the overall population.

*Contact: Ming-Hui Chen, Department of Statistics, University of Connecticut, Storrs, Connecticut 06269, U.S.A. ming-hui.chen@uconn.edu.

The heterogeneity of treatment effects among patients can result from different aspects, including demographical characteristics such as age, gender and marital status; biomedical measurements, which are often referred as bio-markers in practice such as plasma measurements, gene sequencing, etc.; and others. As a consequence, these factors together consist of a high-dimensional covariate space and the selection of subgroup-defining variables is evidently necessary.

While the underlying structure of the treatment effect may never be explained perfectly, several paths have been explored to approximate the truth. The first path is to model the treatment-subgroup interaction as a “black-box” mechanism using penalized regression techniques, as seen in [Zhao et al., 2012], [Tian et al., 2014] and [Zhao et al., 2015]. These model-based approaches are often applied in optimal treatment regime researches, with assigning the prime treatment to each patient as its goal. The second path is to build tree structures to model the treatment effect heterogeneity of different patients, via Classification and Regression Trees (CART)-based [Breiman et al., 1984] recursive partitioning [Su et al., 2009, Dusseldorp and Van Mechelen, 2014], novel unbiased recursive partitioning techniques [Loh et al., 2015], and random forest approaches [Foster et al., 2011]. Various Bayesian methods have also been developed, including Bayesian tree-based methods [Berger et al., 2014, Sivaganesan et al., 2011, Sivaganesan et al., 2017], Bayesian credible subgroup approach [Schnell et al., 2016] and so on. Many of these methods utilize model selection ideas while they have different interpretations comparing to frequentist methods.

Generally, the penalized regression methods produce a complex classifier after performing optimization over a loss/score function, aiming to assign the optimal treatment to each patient. Such black-box mechanisms do not focus a lot on the interpretation of the subgroups and it may not be easy for future patients recruitment in following confirmatory subgroup studies. On the contrary, the tree methods often lead to a partition of the observed patient covariate space, as the terminal nodes of the tree consist of patients whose characteristics meet the conditions of each split, assuming the patients within each subspace have homogeneous treatment effects. It provides practitioners with great convenience in interpretation and implementation. However, these merits will gradually be lost as the tree grows larger and deeper.

While some tree methods [Foster et al., 2011, Sivaganesan et al., 2011, Sivaganesan et al., 2017] provide a decision-rule to choose the final subgroups, most tree methods [Su et al., 2009, Loh et al., 2015, Berger et al., 2014] offer a tree-based patient space partition, without specifying explicit criteria to classify their resulting terminal nodes into beneficial subgroups and non-beneficial groups. Even though the subgroup search process is of exploratory nature, it is still desirable to evaluate and compare how informative these subgroups are under a range of simulated scenarios.

Due to the nature of binary tree structures, it is possible to observe identical trees even though their splitting paths do not match initially. Also, when the true data-generating mechanism is tree-structured, it often ends up with a ‘partially correct’ tree being produced, in which the splitting variable may be correct but the splitting point is wrong, or vice versa. Sometimes the tree is under/over grown comparing to the underlying truth. These issues

impose challenges on the performance evaluation and comparison of different tree methods. An illustrative example is shown in Figure 1, in which the underlying data generating model is $y = 2 + 2I(X_1=0)trt + \epsilon$, where $I(\cdot)$ denotes the 0–1 indicator function.

Various evaluation criteria can be found in the existing literature for performance assessment of tree-based approaches, for instance, checking the final tree size [Su et al., 2009], examining how frequently a non-trivial tree is produced [Dusseldorp and Van Mechelen, 2014], or inspecting how often the true predictive variable is found [Foster et al., 2011]. These criteria do not provide direct information for the performance evaluation of tree-based methods in terms of type I error and power. They also give limited assessment of the tree structure recovery process.

One of our contributions in this paper is to develop evaluation criteria for exploratory tree-based subgroup methods. Specifically, three evaluation criteria are proposed from distinct aspects of the tree restoration process. In addition, we have carried out a comprehensive examination of the empirical performance of several tree-based methods via extensive simulation studies. Advantages and disadvantages of the methods under comparison have been revealed in our simulation studies. These insightful findings along with the proposed evaluation criteria are potentially very useful for both researchers and practitioners in this field.

The rest of the article is structured as follows. Section 2 introduces the general settings and the methods in current tree-based subgroup analysis literature. Section 3 presents the detailed development of the proposed evaluation criteria. In Section 4, we carry out an in-depth investigation of the empirical performance of different tree-based methods under several simulation scenarios. An analysis of the real data is given in Section 5. Section 6 gives brief concluding remarks.

2 Tree-based Subgroup Methods in Comparison

In this section, we introduce general notations to be used throughout the rest of the paper. Consider a trial in which there is a univariate response variable, denoted by Y , with two treatment assignments, placebo denoted by $trt=0$ and the experimental treatment denoted by $trt=1$. Let $\mathbf{X} = (X_1, \dots, X_p)$ be a p -dimensional vector of the patients' baseline covariates. Then the difference between two treatment assignments $\tau(\mathbf{X}) = E(Y|\mathbf{X}, trt=1) - E(Y|\mathbf{X}, trt=0)$ is the primary interest of the study. Without considering the treatment effect heterogeneity among subjects, intrinsically the underlying assumption is $\tau(\mathbf{X}) \equiv c$ for some constant c , which is normally adopted in the drug development process. However, as more and more medical studies have found that the treatment may be more effective on some subject subgroups indicated by certain biomarkers, this assumption may be problematic in practice.

Consider a general nonparametric structure $E(Y|\mathbf{X}, trt) = g(\mathbf{X}, trt)$, where g is a multivariate function representing the underlying mechanism of the signal. Thus $\tau(\mathbf{X}) = E(Y|\mathbf{X}, trt=1) - E(Y|\mathbf{X}, trt=0) = g(\mathbf{X}, 1) - g(\mathbf{X}, 0)$, and we can rewrite the structure equivalently as

$$E(Y | \mathbf{X}, trt) = g(\mathbf{X}, 0) + \Delta(\mathbf{X})trt = a(\mathbf{X}) + \Delta(\mathbf{X})trt. \quad (1)$$

For a normal model with constant variance, it can be expressed as

$$Y = a(X) + \Delta(X)trt + \epsilon, \quad (2)$$

where $\epsilon \sim N(0, \sigma^2)$. In the subgroup analysis literature, the first term $a(X)$ in (1) is often referred as the *prognostic effect*, and $\Delta(X)$ refers to the *predictive effect*. Under the current notation, $a(X)$ can also represent the baseline effects.

While both $a(X)$ and $\Delta(X)$ contribute to the response Y , the estimation of $\Delta(X)$ is of greater importance in the context of subgroup analysis. For tree-based methods, an algorithm will be deployed to firstly grow an initial tree, and prune it back to optimize upon certain criterion. For the non tree-based approaches as often seen in optimal treatment regime studies, a penalized regression scheme will be implemented to estimate $\Delta(X)$ and the patients will be classified into the beneficial subgroup and non-beneficial group at the end of the procedure. Since its estimation process does not model $\Delta(X)$ with an explicit function form, we consider it as a black-box estimation procedure.

In this article, we focus on comparing 4 tree-based methods: Interaction Tree (IT) [Su et al., 2009], Qualitative Interaction Trees (QUINT) [Dusseldorp and Van Mechelen, 2014], Virtual Twins (VT) [Foster et al., 2011] and Generalized Unbiased Interaction Detection and Estimation (GUIDE) [Loh et al., 2015]. These four methods cover a wide spectrum of features among the current exploratory tree-based approaches. For instance, IT, VT, and GUIDE can deal with both qualitative and quantitative interactions in the model, while QUINT is specifically designed to detect only qualitative interaction. IT and QUINT use different criteria motivated by two sample tests to guide the splits, while GUIDE decides split variables using a residual sign test. VT firstly estimates the individual treatment effects via random forest, and then apply recursive partitioning. Additionally, GUIDE does not have selection bias when there are a mixed type of covariates, while IT, VT and QUINT tend to pick the variables which have more categories for splitting than others as reported in [Loh et al., 2015]. Another consideration for selecting these 4 methods for comparison is that all these 4 methods can be implemented either directly or indirectly through R on Windows, Linux or Mac OS X platforms, which makes it possible to carry out extensive simulations on parallel computing clusters. A summary of selected features of these 4 methods is given in Table 1 and more detailed descriptions and discussions of these method are followed in subsequent subsections.

2.1 Interaction Tree (IT) Method

A sequence of papers ([Su et al., 2008], [Su et al., 2009] and [Su et al., 2011]) were published to apply the interaction tree idea towards censored, continuous and binary response variables, respectively. As introduced in the most commonly referred version [Su et al., 2009], the interaction tree method adopts a CART-based recursive partitioning algorithm, guided by the squared t test statistic of the interaction between the treatment and a potential subgroup. Consider the t test statistic $\mathcal{I}(s)$ for testing $H_0: \gamma = 0$ versus $H_1: \gamma \neq 0$ under the model given by

$$y_i = \beta_0 + \beta_1 trt_i + \gamma trt_i z_i^{(s)} + \epsilon, \quad (3)$$

where $z^{(s)} = 1(x_j \in A_j)$ is the indicator associated with split s , and A_j stands for the covariate subspace for x_j that defines the split. Then $G(s) = t(s)^2$ is used as the splitting statistic and the split s^* which attains the maximum $G(s)$ among all possible splits is selected at each splitting step. A large initial tree will be grown at first followed by a pruning procedure using ν -fold cross validation or bootstrapping methods. $G_\lambda(T) = G(T) - \lambda |T - \tilde{T}|$ is the proposed interaction-complexity measure, where $|T - \tilde{T}|$ denotes the number of internal nodes of tree T , $\lambda \geq 0$ is a complexity parameter, and $G(T) = \sum_{h \in T - \tilde{T}} G(h)$ is the overall amount of interaction in tree T . The tree T which achieves the largest $G_\lambda(T)$ will be chosen as the best-sized subtree for a fixed λ . IT provides a terminal node amalgamation procedure which will lead to smaller number of subgroups because groups with homogenous treatment effect will be combined. It also has an algorithm to rank the variable importance.

2.2 Qualitative Interaction Trees (QUINT) Method

QUINT method [Dusseldorp and Van Mechelen, 2014] focuses on the detection of qualitative interaction, where the treatment effect in one subgroup has a different sign comparing to another subgroup. The procedure classifies the patients into 3 classes, namely, $\mathcal{P}_1$, in which the treatment group is better than the placebo group, $\mathcal{P}_2$, in which the treatment group is worse than the placebo group, and $\mathcal{P}_3$, in which there is no difference between the treatment group and the placebo group. It performs recursive binary partitioning of the patient characteristics guided by a global partitioning criterion C , which takes account of both the standardized treatment effect size in each child group and the resulting group size. The first split will be made only if the absolute value of the standardized mean treatment difference in each of the two child nodes exceeds a critical minimum value $d_{\min}$, otherwise the tree will not grow.

Once the first split is made, QUINT uses a stepwise procedure that maximizes the criterion C until reaching some pre-specified stopping criteria, which includes a minimum number of patients per treatment condition and maximum possible number of leaves $L_{upperlimit}$ and so on. A large $d_{\min}$ value tends to produce more trivial trees and thus potentially reduce false tree discoveries, while it may also increase the chance of missing the true qualitative interaction trees in the model. Default values are provided as $d_{\min} = 0.30$ and $L_{upperlimit} = 10$. QUINT performs a bootstrap-based pruning procedure after growing the initial tree. The procedure could be applied to binary, continuous or time-to-event responses.

2.3 Virtual Twins (VT) Method

The virtual twins approach [Foster et al., 2011] aims to identify a covariate subspace A in which the treatment effect is substantially better than the average treatment effect or some prespecified threshold. It would firstly estimate the treatment effect difference W_j for each subject via constructing a random forest using the original data and then make predictions with the treatment assignment switched in a two-arm trial setting. There are two derivatives of the virtual twins method. The first approach is VT(R), which stands for virtual twins

regression and fits a regression tree via R package ‘rpart’ using W_i as the response variable, and all patients are classified into two groups by comparing the prediction of individual treatment effect W_i to the overall treatment effect δ of the whole patient population plus certain threshold. The second approach is virtual twins classification as denoted by VT(C), which will firstly perform a binary classification comparing W_i with δ plus a threshold and then classify the patients again using a classification tree algorithm afterwards.

Originally the VT method was only illustrated in a binary response situation while the idea can be applied in other types of responses in a similar fashion. Without the final classification step in VT(R), the method tends to be more like a tree method, since it offers a tree structured space partition rather than only a dichotomization $\{A, A^c\}$, which is not straightforward for interpretation.

2.4 Generalized Unbiased Interaction Detection and Estimation (GUIDE) Method

GUIDE has been developed and enriched over the past few decades serving for regression [Loh, 2002] and classification [Loh, 2009] purposes primarily. For subgroup analysis, GUIDE uses a regression tree idea [Loh et al., 2015] and applies a two-step procedure for the splitting selection. It performs χ^2 tests on the signs of residuals versus the treatment labels to find the ‘best’ split variable and then searches for the optimal split cutoff. By doing this, GUIDE can eliminate variable selection bias and improve computational efficiency as reported in the original paper.

GUIDE mainly consists of 2 algorithms: Gi (‘i’ for interaction) and Gs (‘s’ for sum). Gi procedure is recommended by the author since it has better performance in handling the prognostic effects than Gs procedure, and we will use ‘GUIDE’ to refer it in the method comparison. Some unique features of GUIDE in comparison with the other methods include that it can deal with missing data naturally and can be applied even when there are more than 2 treatment levels. The procedure can be applied to binary, continuous or censored responses, for both univariate and multivariate cases [Loh et al., 2016].

3 Evaluation Criteria

Most tree-based methods do not label subgroups as “effective” or “ineffective” in terms of the treatment effect. Instead, only a tree partition of the covariate space is presented in the end since the process has more of an exploratory nature. Also, it is not straightforward to apply the conventional type I error and type II error concepts in tree models. For example, it may not be appropriate to define the occurrence of a type I error whenever the tree splits, since it may just produce several terminal nodes but the treatment effect is not significant at certain significance level in every one of them. For the type II error, it would not be very informative to define the occurrence of a type II error as the case that a trivial tree is produced in the end when the true model has treatment subgroup interaction. These measures may have poor discriminative ability because the only information it used is whether the tree splits. In fact, several tree structures may be identical even if their splitting processes are not the same, and the resulted tree may be “partially correct” when it has some common regions shared with the true underlying tree. It is also crucial to compare distinct non-nested trees properly as we may obtain several trees which are all “partially correct”.

There are other measurements in the existing literature, such as counting the average number of terminal nodes and comparing it with the “true” number of terminal nodes in the underlying tree, or examining how frequent the true predictive variables are involved to form the final tree. However, measuring the number of terminal nodes provides no direct information on how well the tree structure is fitted, nor does the matching of variable offers good evidence for tree estimation since it does not consider the splitting cutoffs.

To address these challenging issues, we propose three evaluation criteria: (i) defining false subgroup rate (FSR) to assess the incorrect subgroup detections when there is indeed no subgroup; (ii) constructing the receiver operator characteristic (ROC) curve and using the area under curve (AUC) to evaluate the discrimination ability of each method; (iii) defining a new set of measures T-AIC/T-BIC to evaluate the performance in the sense of structural recovery.

3.1 Criterion 1: False Subgroup Rate (FSR) Evaluation

Assume the final tree has been produced and there are N_T terminal nodes in total. At each terminal node $t=1, \dots, N_T$, we fit the overall treatment model

$$y = \beta_{t0} + \beta_{t1} trt + \eta_t \quad (4)$$

with the patients in the node and get the p -value p_t of the two-sided T-test $H_0: \beta_{t1} = 0$ vs. $H_1: \beta_{t1} \neq 0$. In the s -th simulation repetition, define FSR_s as

$$FSR_s = \begin{cases} 1, & N_T \geq 2 \text{ and } \exists t, s. t. p_t < \alpha; \\ 0, & \text{Otherwise.} \end{cases} \quad (5)$$

and thus the false subgroup rate is defined as $FSR = \sum_s FSR_s / N_S$, where N_S is the total number of repetitions. Notice that if a trivial tree is built, i.e., the tree only contains the root node, we treat it as a case without incorrect subgroup findings. This procedure mimics the fact that people may check each resulted potential subgroup by testing them one by one at a fixed nominal significance level. If an algorithm produces a tree with many terminal nodes, it may have a high FSR value. However, it may not always be true as the partition itself also matters and each test at the terminal node may not be independent. $\alpha = 0.05$ is chosen in our simulations.

3.2 Criterion 2: Receiver Operating Characteristic (ROC) Curve and Area Under Curve (AUC) for Power Study

Since the tree-based method essentially offers a partition of the patient space, we could consider it as a binary classification question similarly to [Schnell et al., 2016] and construct the ROC curve to test its differentiation performance.

Here we define $\mathcal{S}_1$ as the beneficial subgroup, i.e., for which the treatment effect is non-negative, and $\mathcal{S}_0$ is the rest. $\mathcal{S}_1$ shall include all the patients who receive beneficial treatment effects, regardless of their defining characteristics or treatment effect magnitudes. Pre-assign a series of d values, e.g., $\mathcal{D} = \{0, 0.05, 0.1, \dots, 4\}$, serving as the range of threshold values for

classification. The choice of d may be determined by calculating the effect size of each terminal node after applying each tree-based method and, hence, d may be chosen as the absolute value of the largest treatment effect size among all terminal nodes. Fit the overall treatment model (4) to each node t and assign $\hat{d} = \hat{\beta}_{t1}$ to all patients in that node. Then a patient is classified into $\hat{\mathcal{G}}_{1j}$ if $\hat{d}$ is no less than d_j for each $d_j \in \mathcal{D}$.

Define sensitivity (SS) and specificity (SP) as

$$SS = \frac{\# \text{ of patients in } \{\hat{\mathcal{G}}_1 \cap \mathcal{G}_1\}}{\# \text{ of patients in } \mathcal{G}_1}; \quad (6)$$

$$SP = \frac{\# \text{ of patients in } \{\hat{\mathcal{G}}_0 \cap \mathcal{G}_0\}}{\# \text{ of patients in } \mathcal{G}_0}. \quad (7)$$

We average (SS_{js}, SP_{js}) over all simulation repetitions $s=1, \dots, N_S$ and for every $d_j \in \mathcal{D}$, plot $(1 - \bar{S}P_j, \bar{S}S_j)$ across all $d_j \in \mathcal{D}$ to obtain the ROC curve.

The area under curve (AUC) is then used as a measurement of discrimination accuracy. In the ideal case, if a method produces a partition of the entire patient space close to the true underlying structure, the AUC value should be close to 1. However, the AUC measurement cannot account for model complexity of the tree, which may give high ratings for overgrown trees as long as the true underlying tree is nested in them.

3.3 Criterion 3: T-AIC/T-BIC for Structure Recovery

As pointed out previously, the recovery of the predictive effect $\mathbf{X}$ is the main focus within the context, rather than the fitting of the response variable Y . An estimate of $\mathbf{X}$, denoted by $\hat{\Delta}(\mathbf{X})$, is the core contribution from each method and thus it is desirable to measure the dissimilarity of $\mathbf{X}$ and $\hat{\Delta}(\mathbf{X})$ under a proper criterion.

For the final tree produced by each method, it is reasonable to assume that the treatment effect is homogeneous among patients within each terminal node. Otherwise the final tree needs further growth to fully explain the treatment effect heterogeneity, which means more tuning and adjustments are needed beyond the standard product from each method. Hence the fitted predictive effect $\hat{\Delta}(\mathbf{X})$ for a final tree with N_T terminal nodes can be expressed as

$$\hat{\Delta}(\mathbf{X}) = \sum_{t=1}^{N_T} \gamma_t I(\mathbf{X} \in \mathcal{X}_t) = \mathbf{Z}\boldsymbol{\gamma}, \quad (8)$$

where $I(\mathbf{X} \in \mathcal{X}_t)$ is the indicator function of the conditions that define terminal node t , and $\boldsymbol{\gamma} = (\gamma_1, \dots, \gamma_{N_T})$ is the $N_T \times 1$ vector of constant treatment effects for patients in each node $t=1, \dots, N_T$. $\mathbf{Z}$ is the $n \times N_T$ terminal node assignment matrix such that $Z_{ij}=1$ if the i -th individual is allocated in terminal node j and $Z_{ij'}=0$ for all $j' \neq j, j'=1, \dots, N_T$. In Figure 2,

we show four examples of the terminal node assignment matrices next to each tree diagram, the number of columns of each Z is equal to the number of terminal nodes of the corresponding tree.

Since the sum of the column vectors in Z equals $\mathbf{1}$, we fit the linear regression model $(X) \sim Z$ without the intercept, and obtain its residual sum of squares RSS . In relation to the well-known information criteria AIC and BIC, we define T-AIC and T-BIC (T stands for tree) as

$$T - AIC = n \log[\max(RSS, dt)] + 2N_T; \quad (9)$$

$$T - BIC = n \log[\max(RSS, dt)] + \log(n)N_T, \quad (10)$$

where dt is the *differential threshold* parameter, which pre-specifies the “desired” minimum difference of “goodness of fit”. The model complexity penalty component in the T-AIC/T-BIC overcomes the drawback of AUC and penalizes overgrown trees.

In the current linear regression setting, when dt approaches 0, the proposed T-AIC/T-BIC measure will differ from the original AIC/BIC only by a constant related to sample size n . If the values of RSS for several trees structures are all less than the threshold dt , we treat them as the same and promote the simplest tree among them. If RSS of a tree is larger than dt , the measure also reduces to the original AIC/BIC with a constant difference. The inclusion of dt could be practically meaningful and facilitates implementation in accordance with the spirit of Occam’s razor.

Notice that in the case when there is no treatment effect in the model, i.e., $(X) \equiv 0$, the RSS of the linear regression is always 0, hence T-AIC/T-BIC reduces to $n \log(dt) + 2N_T$ and $n \log(dt) + \log(n)N_T$, respectively. The preference will be given to trees with less terminal nodes and for a fixed n , the smallest value is obtained when N_T , i.e., the tree only contains the root node and the treatment effect is homogeneous in the entire population. Thus the proposed T-AIC/T-BIC measure is applicable and also meaningful in false subgroup rate studies.

When the underlying predictive effect structure (X) is not a tree, the T-AIC/T-BIC criteria will evaluate the performance of approximation provided by the fitted tree model $\hat{\Delta}(X)$ and account for the model complexity simultaneously.

In the T-AIC/T-BIC criterion, the true underlying treatment effect structure (X) is assumed known and it can be applied when comparing different subgroup methods in simulation study. When analyzing with real data where (X) is unknown, an empirical version framework is developed to serve as an approximation to the original measure. Considering a balanced treatment design, mimicking the modified covariate idea in [Tian et al., 2014], we recode the treatment term trt into ± 1 by letting $trt^* = 2trt - 1$, plug it into formula (1) and get

$$E(Y | X, trt^*) = a(X) + \frac{\Delta(X)trt^* + \Delta(X)}{2}, \quad (11)$$

which can be also written as $E(\frac{2Y}{trt^*} | X, trt^*) = \frac{2a(X) + \Delta(X)}{trt^*} + \Delta(X)$. For a balanced treatment design, $E(\frac{2a(X) + \Delta(X)}{trt^*}) = 0$ since $E(\frac{1}{trt^*}) = 0$ and trt^* is independent of X .

Define the modified response $Y^* = \frac{2Y}{trt^*}$. We then fit $Y^* \sim (X)$ by implementing a nonparametric regression method, for instance, the R package ‘rpart’, with all X as potential predictors. Let $\tilde{\Delta}(X)$ equal to the predicted values of Y^* , and we define the empirical version of T-AIC/T-BIC via treating $\tilde{\Delta}(X)$ as the true (X) when computing RSS in equations (9) and (10). For any unbalanced two-arm treatment design, similarly we can recode the treatment term to make $E(\frac{1}{trt^*}) = 0$ and obtain $\tilde{\Delta}(X)$ in the same fashion.

In simulation studies, we used R package ‘rpart’ for the regression of Y^* with X and found that when sample size is large comparing to the number of the covariates, this empirical version of T-AIC/T-BIC provides nearly identical rankings of model selection percentages comparing to the original T-AIC/T-BIC using the true (X) . Several examples are included in Section 4.3.

4 Simulation Studies

In this section, we evaluate the performance of each method using the proposed criteria. Here we refer to the false subgroup rate study as the scenarios in which the treatment effect is 0 for all patients, and refer to the power study as the scenarios in which the treatment effect is non-zero for at least part of the patient population. We design and present several simulations for each of the false subgroup rate study and the power study, aiming to examine the performance of each method corresponding to various setups of effect structure, sample size and correlation among variables.

The softwares are available at <http://biopharmnet.com/subgroup-analysis-software/>. All of the analyses of the simulated data were conducted using R. We used the default choices for the tuning parameters if available. For the IT method, we used the complexity parameter $\lambda = \log(n)$ as suggested in [Su et al., 2009], and applied the bootstrap procedure with 25 bootstrap samples to select the best tree size when pruning. For the QUINT method, we used the R package ‘quint’, set the critical minimum value $d_{\min} = 0.3$ and used 25 bootstrap samples in the pruning procedure. For the Virtual Twins we applied the ‘randomForest’ package with number of trees equal to 1000 when computing the individual treatment contrast, and used ‘rpart’ package with default settings to build up the regression tree. We didn’t perform the final classification step as mentioned in the VT(R) procedure, and the regression tree structure would be treated as the final outcome. For the GUIDE method, we downloaded the latest GUIDE software written by Fortran from <http://>

www.stat.wisc.edu/loh/guide.html. The user manual of GUIDE is also available there and the standard procedures are called from R when implementing Gi procedure.

Throughout all simulation studies we simulate the covariates as follows. Firstly we generate the treatment variable $trt \sim \text{Bernoulli}(0.5)$, covariates $X_1 \sim \text{Bernoulli}(0.5)$, $X_2 \sim \mathcal{N}(0.5, 2^2)$, and $X_3 \sim \text{Bernoulli}(0.7)$. Then we simulate $p-3$ normal covariates X_4 to X_p from $\mathcal{N}(\mathbf{0}_{p-3}, R)$, where p is the total number of covariates. Assign a first-order autoregressive correlation structure to $X_4 - X_p$, with covariance matrix $R_{ij} = \rho^{|i-j|}$. The trt , X_1 , X_2 , X_3 and $\{X_4, \dots, X_p\}$ are mutually independent. Notice that continuous covariates have more possible splits comparing to binary covariates, and some methods may have selection bias as discussed and compared in GUIDE's article [Loh et al., 2015].

The total sample size n is selected at 100, 300 or 500, representing small to large sample size setups. The total number of candidate variables p is tested at 10, 50 separately in order to check the effects of 'noise' variables which play no role in the treatment effect structure. The random noise level σ^2 defined in normal model (2) is tested at $\sigma=1, 2$ separately. We generate the response Y based on the true model in each scenario, which will be introduced under each section. In each simulation, the $n \times (p+2)$ data matrix $(Y, trt, X_1, \dots, X_p)$ will be input to construct the tree for each method.

In each simulation, the method(s) which produced the smallest T-AIC/T-BIC value will be "selected". If there are more than one method with same T-AIC/T-BIC values, all of them will be "selected". Define the method selection percentage (MSP) of method i by

$$MSP_i = \frac{\text{\#of selections of method } i}{\text{\#of selections of all methods}}, \text{ and we have } \sum_i MSP_i = 1.$$

In our false subgroup rate study, $(X) \equiv 0$. The comparison of T-AIC/T-BIC will reduce to the comparison of the number of terminal nodes. We define TML_{mean} as the mean of the number of terminal nodes.

4.1 False Subgroup Rate Study

For the false subgroup rate analysis we test on the following four models: [(a)]

$$(a) y = 2 + \epsilon;$$

$$(b) y = 2 + 2X_4 + \epsilon.$$

$$(c) y = 2 + 2I(X_1 = 0) + 2I(X_2 > 0) + \epsilon;$$

$$(d) y = 2 + 2[I(X_1 = 0) + I(X_2 > 0) + \exp(X_4) + (X_5 + X_6)^2] + \epsilon;$$

Here model (a) represents the simplest situation that only an intercept term serves as the signal. In model (b) we have one linear prognostic variable, and GUIDE claims to control the “type I error” well in this case. In model (c), we designed a tree-structured prognostic effect with 2 prognostic variables. In model (d), non-linear and non-tree structured prognostic effect terms are added to present a complicated prognostic effect formation. The prognostic effect structure in model (c) is the same as adopted in the following power study. Each simulation setup is evaluated based on 1000 repetitions.

To compare the performance of each method side-by-side, in Table 2, we present the FSR values and method selection percentages (MSP) by T-BIC across different settings of n , p for model (a)-(d) with $\rho=0$, $\sigma=1$ and we present the rest of the simulation results in Table A.5 and Table A.7 in the Appendix Section. Only the method selection percentages produced by T-BIC are shown since they are very close to those produced by T-AIC.

From Table 2, overall IT and GUIDE have noticeably lower FSR comparing to QUINT and VT. In particular, in model (b) which has one linear prognostic variable, FSR of GUIDE is less than 0.005, which validates the claim that GUIDIE control FSR pretty well when there is only one prognostic variable.

When sample size n increases while keeping other parameters fixed, we can see from Table 2, that the FSR decreases for IT, QT, but it increases for VT across all models (a)-(d). For example, in model setting (a) with $p=10$, $\sigma=1$, $\rho=0$, when n is 100, 300 and 500, respectively, the FSR of VT is 0.23, 0.55 and 0.726 correspondingly. When we looked back into each simulation repetition, we found that the tree produced by VT would grow some small-sized branch, with significant overall treatment effect at level 0.05. When sample size increases, it is more likely there exists such small subgroups with significant overall treatment effect due to the random noise. Regarding to GUIDE, when there is no more than 1 prognostic variable, from model (a) and (b) we can see FSR decreases as sample size increases; When there is more than 1 prognostic variable, from results of model (c) and (d) we observe that FSR for GUIDE increases as sample size increases.

When the number of variable p increases from 10 to 50, we observe that the FSR of IT and GUIDE decreases overall, and FSR of QUINT and VT increases. When the standard deviation of the random error σ increases from 1 to 2, there is no noticeable difference for all methods, which is mainly due to the fact that we use t-test in the examination procedure.

In terms of the MSP measure, the preferences among these methods induced by the MSP measure are largely consistent with the findings by FSR measure. Also, we can see more distinctive patterns across each method when considering MSP measure. For instance, for model (d) in Table 2, the MSPs for VT are almost 0 for all setups of n and p , and the FSR values are high as well. When sample size increases to 500, the MSP of QUINT is 0.398 when $p=10$ and 0.322 when $p=50$, which demonstrates that QUINT is preferable to GUIDE and is comparable with IT in those cases.

Overall, IT and GUIDE are selected the most often, then followed by QUINT. VT is barely selected across all scenarios, which may be due to its over-splitting issue. As a advantage comparing to FSR measure, if certain method produces a non-trivial tree in which none of

the branches has significant overall treatment effect, it is treated as no false subgroup finding presents in the FSR measure computation, while MSP can account for its structural difference comparing to the true predictive effect structure, i.e., $(\mathbf{X})$.

4.2 Power Analyses

For the power analysis we test on three major types of models: [(i)]

$$(i) y = 2 + c_0[I(X_1 = 0) + I(X_2 > 0)] + 2I(X_4 > 0)trt + \epsilon;$$

$$(ii) y = 2 + c_0[I(X_1 = 0) + I(X_2 > 0)] + 2[I(X_3 = 0) + I(X_4 > 0)]trt + \epsilon;$$

$$(iii) y = 2 + c_0[I(X_1 = 0) + I(X_2 > 0)] + 2(X_5 + X_6)trt + \epsilon,$$

c_0 takes 0 or 2 in correspondence to scenarios without or with prognostic effects, respectively, and thus we have 6 data generating models in total. We would expect that the presence of prognostic effect will negatively affect the performance of each method, especially when there are more than 1 prognostic variables and the size of effect is comparable to the predictive effects.

In model (i) we only have one predictive variable and ideally only one split on X_4 at 0 will be necessary to construct the tree. In model (ii), two independent predictive variables are present and the ideal tree would split on X_3 at 0 and split on X_4 at 0 for both branches produced by the first split (or reversed the order of splitting variables). As often seen in large trees, some terminal nodes in the same tree may have identical treatment effect size even though they come from distinct branches, thus the AUC measure is not as informative as the T-AIC/T-BIC measure since it only considers the binary classification performance in the end without evaluating the goodness of structure recovery. 200 independent simulations are conducted for each scenario.

We present the results of average AUC values and MSP derived by T-BIC, across different settings of n, p for model (i)-(iii) with $\rho=0, \sigma=1$ in Table 3. The rest of the results are presented in Tables A.8 and A.10 within Appendix Section. From Table 3 we can see that VT has the highest AUC values among these 4 methods in general. In model (i) and (ii), the predictive effect are tree-structured, and there is no qualitative treatment subgroup interaction, thus we can see the AUC of QUINT decreases as sample size n increases. In model (iii), though the predictive effect is not tree-structured, there exists a qualitative treatment subgroup interaction, that is, $2(X_5+X_6)trt$. In this case AUC is not monotonically decreasing when n increases.

When sample size n increases, average AUC values of IT, VT and GUIDE increase overall as expected. For model (i) with $c_0 = 0, n = 500, \rho = 0, \sigma = 1$ as shown in Table 3, the AUC values of IT, VT and GUIDE are all about 0.996, which indicates the binary classification of patients are very accurate. Since ideally model (i) should have only two branches, i.e., two

subgroups, those results also imply that the structural recovery performances of those 3 methods are likely close. This can be seen in terms of MSP. For instance, when $n=500$, $p=10$, the MSPs of IT, VT and GUIDE are 0.303, 0.253, 0.434, respectively, which are close. However, as mentioned previously, AUC and MSP will not always act in the same direction because AUC does not account for the dimension of the tree. Certain methods may produce overgrown trees, such that the AUC value is large, but its T-BIC value will also be large and the tree will not be preferred in terms of MSP.

When changing c_0 from 0 to 2, i.e., adding two prognostic variables into each model, the results of QUINT in terms of AUC change the least among these 4 methods, and IT seems to be the least robust method in the group especially when the sample size is as small as 100.

When random noise level σ increases from 1 to 2 as seen in Tables 3 and A.8, the AUCs of all methods have noticeable decrease. The same trend is observed when the number of candidate variables p increases from 10 to 50, which implies there are more “decoys” for each method since there are only a few true predictive variables in each of models (i)-(iii).

In terms of the MSP measure, we can see from these tables that VT would still be the best overall, especially when dealing with the non-tree structured interaction effect in model (iii). When sample size is moderate or large, GUIDE method is comparable or even better than VT in certain cases. For example, in Table 3 MSP of VT is 0.253 in model (i) with $c_0 = 0$, $n = 500$, $p = 10$, $\rho = 0$, $\sigma = 1$, while the MSP for GUIDE is 0.434 in that case. The QUINT method had more selections in certain cases such as model (ii) with $c_0 = 2$, $n = 100$, $p = 50$, $\rho = 0$, $\sigma = 1$ in Table 3, partially because it may produce smaller, but partially correct trees in those setups comparing to other methods. However, a potential reason why QUINT is not even the 2nd best method for model (iii) with a qualitative interaction is that the qualitative interaction effect is non-tree structured, and QUINT may classify certain patients into its $\mathcal{P}_3$ group, in which there is no difference between the treatment and placebo groups.

4.3 Empirical Measure Simulation Examples

We use the same variable generating mechanism as described previously, the sample size n is tested at 3 levels: 200, 600 and 1000. The number of candidate variables p is set at 10 and 30 to represent small and moderate dimension of covariate space. The noise level $\sigma=1$. We use R package ‘rpart’ to fit the regression between modified response Y^* and all covariates X , and average the results over 200 independent simulations for each parameter setup.

4.3.1 False Subgroup Rate Study Simulation Example—In the false subgroup rate study example we let the underlying true model be: $y=2+2I(X_1=0)+2I(X_2>0)+\epsilon$ with no correlation among candidate predictors X . The results using the T-BIC measures are displayed in Figure 3, we can see that when sample size is large (more than 600) comparing to number of variables, the rankings of the MSPs provided by the original T-BIC and the empirical T-BIC are almost identical.

4.3.2 Power Study Simulation Example—In the power study example we let the underlying true model be: $y=2+2I(X_4>0)trt+\epsilon$ with no correlation among candidate predictors X . The results using the T-BIC measures are displayed in Figure 4, we can see

that when sample size n is large (more than 600) comparing to number of variables p , the rankings of the MSPs provided by the original T-BIC and the empirical T-BIC are almost identical. When $p=30$, more samples are needed to approximate the true T-BIC accurately comparing to the cases when $p=10$.

5 A Real Data Application

To illustrate the need of method comparison, we consider the Breast Cancer Recovery Project (BCRP) dataset as an example to inspect the results produced by the 4 tree-based methods in comparison. The dataset is about women with early-stage breast cancer, and was first analyzed in [Scheier et al., 2007]. There are three arms in the randomized trial: a nutrition intervention arm with 85 patients, an education intervention arm with 83 patients, and a standard care arm with 84 patients. In our analysis, we only use the patients from the education intervention arm and the standard care arm. Assign $trt=1$ for the patients who received education intervention, and $trt=0$ for the patients with standard care. After removing missing values, we have in total $n=146$ subjects, among which 70 patients are in $trt=1$ group and 76 patients are in $trt=0$ group.

The BCRP dataset contains the baseline measurement and measurement from a 9-month follow-up. The response variable is the improvement in depression score, and the same 9 covariates are used in the analysis as adopted in [Dusseldorp and Van Mechelen, 2014]: age, nationality, marital status, weight change, treatment extensiveness index, comorbidities, dispositional optimism, unmitigated communion and negative social interaction. A brief summary for all the variables involved here is displayed in Table 4. More details of the dataset can be found in [Scheier et al., 2007] and [Dusseldorp and Van Mechelen, 2014].

Here we apply the previously discussed 4 tree-based subgroup methods to the BCRP dataset, and the trees produced by these methods are displayed in Figure 5. The sample size in each terminal node is labeled below each terminal itself. The empirical T-AIC/T-BIC values are calculated and for simplicity we only discuss the empirical T-BIC results here. Firstly we can see from Figure 5 that the IT and GUIDE methods produced a trivial tree, which only had the root node. The QUINT and VT methods produced two different trees with the same number of terminal nodes. In particular, the tree produced by QUINT is grown upon the estimated structure $\tilde{\Delta}(X)$, which would perfectly explain the heterogeneity when considering $\tilde{\Delta}(X)$ to be the truth, and thus QUINT had the smallest empirical T-BIC value comparing with those of IT, GUIDE and VT.

6 Concluding Remarks

In this article several tree-based methods are compared using different proposed criteria. The FSR criterion is defined in connection with the conventional type I error rate concept by mimicking the naive idea of “test them all” at all terminal nodes. The AUC measure is defined by treating each tree-based method as a binary classifier and serves as a measurement of power. It does not account for the model complexity and may not promote simpler trees.

The T-AIC/T-BIC measures consider both goodness-of-fit and model complexity simultaneously and achieves a delicate balance to select the “optimal” structure by mimicking the classic AIC/BIC. Between these two measures, T-BIC is more preferred as it avoids over complicated models when their goodness of fit are similar to certain degree controlled by dt , which is a desirable feature in practice. The choice of dt can be guided by the ‘scale’ of the response variable such as standard deviation and pre-specified. It could be implemented in both false subgroup rate study and power study since we have no restriction on the underlying truth $b(X)$ in simulation studies. Moreover, the T-AIC/T-BIC measure provides more distinctive patterns across each method and facilitates the decision making process.

Based on the simulation results, VT has issues with high FSR which increases with larger sample size though it also has higher AUC when sample size increases, so VT may not be considered if the FSR is of great concern. QUINT method is specifically designed for scenarios with qualitative interaction, but in the simulation scenario with a qualitative interaction QUINT was not even the 2nd best method if we consider AUC, and it was outperformed by IT and GUIDE in terms of FSR when at most 1 prognostic variable is present in the model. IT is better in terms of FSR and GUIDE is better in terms of power comparing to other approaches. However, overall GUIDE is a good approach if we have to pick one of these 4 approaches.

Certainly, our simulations may be limited since the underlying truth may be intricate and thus may never be perfectly revealed. Still we wish to use these representative examples to illustrate how each method behaves under our proposed criteria, and provide some general ideas for future users when the method selection is needed.

In practice, the empirical version of the T-AIC/T-BIC may be used as a reference when the sample size n comparing to the number of covariates, p , is large. Theoretical arguments for the T-AIC/T-BIC and its empirical version are very desirable and to be constructed in the future work.

Acknowledgements

We would like to thank the Editor in Chief, the Associate Editor, and the two referees for their helpful and constructive comments and suggestions, which have led to a much improved version of the manuscript. We are also grateful to Drs. Elise Dusseldorp, Jared Foster, Ilya Lipkovich, Wei-Yin Loh, and Xiaogang Su for making the softwares of the compared tree-based methods publicly available on <http://biopharmnet.com/subgroup-analysis-software/>, and for their patience as well as constructive comments. Mr. Liu’s research was partially supported by NIH grant #P01CA142538. Dr. Chen’s research was partially supported by NIH grants #GM70335 and #P01CA142538.

Appendix: Additional Tables for Simulation Study

Appendix A.1: False Subgroup Rate Study Results

Table A.5:

FSR values and MSP values for models (a)-(d) with $\rho=0$, $\sigma=2$.

			IT		QUINT		VT		GUIDE	
Model	Measure	np	10	50	10	50	10	50	10	50
(a)	FSR	100	0.067	0.025	0.583	0.745	0.246	0.664	0.004	0.011
		300	0.000	0.000	0.176	0.374	0.562	0.886	0.007	0.003
		500	0.000	0.000	0.025	0.100	0.726	0.626	0.003	0.006
	MSP	100	0.409	0.440	0.156	0.115	0.000	0.000	0.435	0.445
		300	0.355	0.381	0.293	0.239	0.000	0.000	0.353	0.380
		500	0.336	0.345	0.328	0.310	0.000	0.002	0.335	0.343
(b)	FSR	100	0.078	0.025	0.595	0.740	0.224	0.531	0.003	0.002
		300	0.001	0.000	0.178	0.386	0.520	0.678	0.002	0.000
		500	0.000	0.000	0.027	0.100	0.666	0.390	0.002	0.001
	MSP	100	0.408	0.437	0.153	0.116	0.000	0.000	0.440	0.447
		300	0.355	0.383	0.292	0.235	0.000	0.000	0.354	0.383
		500	0.337	0.345	0.327	0.310	0.000	0.000	0.336	0.345
(c)	FSR	100	0.078	0.026	0.568	0.734	0.206	0.658	0.024	0.012
		300	0.000	0.000	0.184	0.397	0.555	0.718	0.114	0.038
		500	0.000	0.000	0.027	0.098	0.709	0.368	0.199	0.036
	MSP	100	0.410	0.438	0.168	0.119	0.000	0.000	0.422	0.443
		300	0.395	0.393	0.322	0.237	0.000	0.000	0.283	0.370
		500	0.392	0.350	0.381	0.316	0.000	0.003	0.227	0.330
(d)	FSR	100	0.115	0.044	0.580	0.751	0.161	0.416	0.022	0.006
		300	0.011	0.002	0.155	0.342	0.428	0.648	0.114	0.019
		500	0.010	0.002	0.026	0.087	0.497	0.413	0.202	0.061
	MSP	100	0.405	0.435	0.165	0.114	0.004	0.001	0.427	0.451
		300	0.388	0.382	0.334	0.252	0.000	0.000	0.277	0.366
		500	0.392	0.355	0.390	0.325	0.000	0.001	0.218	0.319

Table A.6:

FSR values and MSP values for models (a)-(d) with $\rho=0.5$, $\sigma=1$.

			IT		QUINT		VT		GUIDE	
Model	Measure	np	10	50	10	50	10	50	10	50
(a)	FSR	100	0.079	0.042	0.568	0.749	0.259	0.469	0.015	0.013
		300	0.000	0.000	0.167	0.385	0.550	0.811	0.001	0.003
		500	0.000	0.000	0.019	0.079	0.744	0.657	0.001	0.004
	MSP	100	0.404	0.436	0.164	0.114	0.000	0.000	0.432	0.449

			IT		QUINT		VT		GUIDE	
Model	Measure	np	10	50	10	50	10	50	10	50
		300	0.353	0.383	0.294	0.235	0.000	0.000	0.353	0.382
		500	0.336	0.343	0.329	0.316	0.000	0.000	0.335	0.341
(b)	FSR	100	0.081	0.039	0.529	0.735	0.196	0.376	0.002	0.004
		300	0.000	0.000	0.148	0.387	0.483	0.525	0.001	0.001
		500	0.000	0.000	0.022	0.088	0.607	0.433	0.000	0.001
	MSP	100	0.396	0.433	0.175	0.119	0.000	0.000	0.429	0.448
		300	0.351	0.383	0.299	0.235	0.000	0.000	0.350	0.382
		500	0.336	0.344	0.328	0.313	0.000	0.000	0.336	0.343
(c)	FSR	100	0.098	0.034	0.557	0.745	0.202	0.413	0.109	0.020
		300	0.000	0.000	0.164	0.386	0.482	0.506	0.345	0.090
		500	0.000	0.000	0.033	0.102	0.623	0.497	0.508	0.138
	MSP	100	0.445	0.443	0.192	0.115	0.005	0.000	0.359	0.441
		300	0.467	0.407	0.390	0.250	0.000	0.000	0.142	0.343
		500	0.476	0.375	0.460	0.337	0.000	0.000	0.065	0.288
(d)	FSR	100	0.154	0.046	0.508	0.748	0.155	0.442	0.053	0.012
		300	0.027	0.002	0.150	0.356	0.358	0.470	0.262	0.045
		500	0.020	0.004	0.021	0.091	0.390	0.353	0.360	0.109
	MSP	100	0.399	0.437	0.198	0.115	0.004	0.000	0.399	0.449
		300	0.435	0.393	0.383	0.253	0.000	0.000	0.182	0.354
		500	0.445	0.366	0.449	0.334	0.001	0.000	0.105	0.300

Table A.7:

FSR values and MSP values for models (a)-(d) with $\rho=0.5$, $\sigma=2$.

			IT		QUINT		VT		GUIDE	
Model	Measure	np	10	50	10	50	10	50	10	50
(a)	FSR	100	0.075	0.029	0.568	0.738	0.256	0.468	0.012	0.008
		300	0.000	0.000	0.182	0.391	0.584	0.800	0.005	0.003
		500	0.000	0.000	0.025	0.093	0.736	0.667	0.003	0.001
	MSP	100	0.403	0.436	0.167	0.118	0.000	0.000	0.430	0.446
		300	0.356	0.384	0.291	0.234	0.000	0.000	0.353	0.382
		500	0.336	0.344	0.328	0.312	0.000	0.000	0.335	0.344
(b)	FSR	100	0.085	0.033	0.545	0.730	0.195	0.440	0.000	0.007
		300	0.002	0.000	0.159	0.374	0.497	0.647	0.001	0.006
		500	0.000	0.000	0.033	0.112	0.680	0.409	0.001	0.002
	MSP	100	0.397	0.434	0.169	0.121	0.000	0.000	0.434	0.445
		300	0.352	0.382	0.296	0.239	0.000	0.000	0.352	0.379
		500	0.337	0.347	0.326	0.308	0.000	0.000	0.337	0.346

			IT		QUINT		VT		GUIDE	
Model	Measure	np	10	50	10	50	10	50	10	50
(c)	FSR	100	0.068	0.028	0.551	0.730	0.197	0.525	0.022	0.007
		300	0.000	0.000	0.189	0.363	0.558	0.643	0.129	0.032
		500	0.000	0.000	0.033	0.086	0.690	0.451	0.199	0.064
	MSP	100	0.412	0.435	0.172	0.121	0.000	0.000	0.416	0.444
		300	0.396	0.388	0.320	0.247	0.000	0.000	0.284	0.366
		500	0.392	0.354	0.379	0.324	0.000	0.000	0.229	0.322
(d)	FSR	100	0.162	0.036	0.544	0.748	0.189	0.486	0.038	0.009
		300	0.018	0.006	0.130	0.393	0.355	0.562	0.246	0.044
		500	0.018	0.002	0.017	0.082	0.401	0.385	0.366	0.091
	MSP	100	0.404	0.439	0.183	0.112	0.003	0.000	0.409	0.449
		300	0.424	0.393	0.378	0.240	0.000	0.000	0.198	0.367
		500	0.439	0.361	0.445	0.332	0.001	0.000	0.115	0.306

Appendix A.2: Power Study Results

Table A.8:

AUC values and MSP values for models (i)-(iii) with $\rho=0$, $\sigma=2$.

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
(i)	0	AUC	100	0.547	0.511	0.652	0.604	0.877	0.694	0.542	0.535
			300	0.524	0.500	0.572	0.606	0.955	0.972	0.820	0.758
			500	0.603	0.500	0.544	0.541	0.980	0.986	0.953	0.904
		MSP	100	0.104	0.191	0.173	0.181	0.678	0.451	0.045	0.176
			300	0.005	0.000	0.010	0.035	0.750	0.785	0.235	0.180
			500	0.020	0.000	0.005	0.010	0.587	0.730	0.388	0.260
	2	AUC	100	0.555	0.506	0.676	0.589	0.764	0.559	0.533	0.509
			300	0.505	0.500	0.599	0.606	0.932	0.892	0.845	0.745
			500	0.519	0.500	0.539	0.539	0.970	0.975	0.974	0.954
		MSP	100	0.184	0.301	0.291	0.187	0.432	0.177	0.092	0.335
			300	0.000	0.010	0.040	0.065	0.680	0.705	0.280	0.220
			500	0.005	0.000	0.005	0.000	0.455	0.505	0.535	0.495
(ii)	0	AUC	100	0.560	0.509	0.679	0.612	0.767	0.686	0.572	0.531
			300	0.547	0.500	0.598	0.598	0.911	0.866	0.813	0.723
			500	0.688	0.506	0.520	0.544	0.976	0.959	0.929	0.879
		MSP	100	0.030	0.005	0.745	0.805	0.205	0.175	0.020	0.015
			300	0.000	0.000	0.130	0.250	0.735	0.655	0.135	0.095
			500	0.005	0.005	0.000	0.055	0.890	0.815	0.105	0.125
	2	AUC	100	0.550	0.510	0.645	0.591	0.701	0.600	0.541	0.525

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
			300	0.507	0.500	0.588	0.581	0.861	0.810	0.764	0.663
			500	0.536	0.502	0.539	0.549	0.934	0.882	0.921	0.858
		MSP	100	0.035	0.000	0.780	0.905	0.160	0.075	0.025	0.020
			300	0.000	0.000	0.340	0.595	0.470	0.260	0.190	0.145
			500	0.000	0.000	0.055	0.155	0.620	0.495	0.325	0.350
(iii)	0	AUC	100	0.711	0.520	0.725	0.706	0.807	0.772	0.650	0.582
			300	0.793	0.650	0.797	0.779	0.883	0.868	0.798	0.773
			500	0.825	0.763	0.778	0.771	0.885	0.878	0.831	0.813
		MSP	100	0.370	0.040	0.130	0.235	0.440	0.655	0.060	0.070
			300	0.000	0.000	0.015	0.050	0.965	0.950	0.020	0.000
			500	0.010	0.000	0.025	0.025	0.965	0.970	0.000	0.005
	2	AUC	100	0.647	0.516	0.724	0.692	0.779	0.710	0.628	0.562
			300	0.719	0.540	0.796	0.773	0.864	0.853	0.792	0.739
			500	0.796	0.705	0.791	0.785	0.877	0.857	0.832	0.810
		MSP	100	0.170	0.050	0.235	0.245	0.525	0.610	0.070	0.095
			300	0.000	0.000	0.065	0.080	0.930	0.910	0.005	0.010
			500	0.000	0.000	0.015	0.050	0.970	0.940	0.015	0.010

Table A.9:

AUC values and MSP values for models (i)-(iii) with $\rho=0.5$, $\sigma=1$.

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
(i)	0	AUC	100	0.889	0.557	0.758	0.691	0.977	0.975	0.897	0.811
			300	0.991	0.953	0.571	0.600	0.992	0.991	0.993	0.995
			500	0.994	0.995	0.505	0.529	0.996	0.996	0.996	0.996
		MSP	100	0.241	0.034	0.093	0.073	0.426	0.678	0.241	0.215
			300	0.302	0.249	0.014	0.023	0.222	0.249	0.462	0.480
			500	0.323	0.293	0.000	0.000	0.262	0.314	0.415	0.393
	2	AUC	100	0.578	0.515	0.718	0.696	0.898	0.740	0.730	0.598
			300	0.599	0.500	0.612	0.587	0.989	0.985	0.988	0.984
			500	0.797	0.547	0.547	0.543	0.992	0.993	0.992	0.994
		MSP	100	0.060	0.126	0.190	0.320	0.545	0.364	0.205	0.189
			300	0.035	0.000	0.010	0.020	0.448	0.395	0.507	0.585
			500	0.138	0.025	0.015	0.020	0.345	0.350	0.502	0.605
(ii)	0	AUC	100	0.838	0.553	0.721	0.727	0.873	0.848	0.836	0.755
			300	0.974	0.903	0.579	0.580	0.994	0.990	0.993	0.991
			500	0.991	0.976	0.534	0.524	0.996	0.997	0.996	0.996

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
		MSP	100	0.340	0.030	0.335	0.550	0.180	0.285	0.145	0.135
			300	0.069	0.005	0.025	0.040	0.685	0.703	0.222	0.252
			500	0.184	0.025	0.024	0.015	0.578	0.750	0.214	0.210
	2	AUC	100	0.590	0.525	0.697	0.686	0.785	0.683	0.692	0.585
			300	0.633	0.511	0.613	0.630	0.962	0.884	0.966	0.928
			500	0.837	0.603	0.572	0.555	0.993	0.972	0.989	0.991
		MSP	100	0.035	0.005	0.695	0.855	0.155	0.085	0.115	0.055
			300	0.005	0.000	0.080	0.190	0.535	0.270	0.380	0.540
			500	0.000	0.000	0.030	0.045	0.740	0.465	0.230	0.490
(iii)	0	AUC	100	0.802	0.590	0.819	0.814	0.882	0.880	0.865	0.849
			300	0.914	0.876	0.853	0.840	0.915	0.913	0.904	0.890
			500	0.925	0.895	0.857	0.846	0.918	0.912	0.921	0.903
		MSP	100	0.530	0.110	0.145	0.090	0.145	0.545	0.180	0.255
			300	0.375	0.055	0.030	0.065	0.400	0.740	0.195	0.140
			500	0.395	0.055	0.000	0.005	0.185	0.610	0.420	0.330
	2	AUC	100	0.792	0.557	0.809	0.809	0.866	0.875	0.814	0.763
			300	0.870	0.846	0.849	0.844	0.902	0.905	0.884	0.876
			500	0.876	0.866	0.849	0.840	0.912	0.907	0.907	0.888
		MSP	100	0.415	0.030	0.120	0.145	0.310	0.705	0.155	0.120
			300	0.020	0.000	0.020	0.035	0.810	0.895	0.150	0.070
			500	0.000	0.000	0.000	0.025	0.625	0.855	0.375	0.120

Table A.10:

AUC values and MSP values for models (i)-(iii) with $\rho=0.5$, $\sigma=2$.

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
(i)	0	AUC	100	0.561	0.508	0.632	0.627	0.844	0.714	0.542	0.525
			300	0.538	0.500	0.583	0.614	0.952	0.966	0.844	0.787
			500	0.619	0.505	0.534	0.539	0.978	0.987	0.964	0.918
		MSP	100	0.113	0.162	0.143	0.201	0.680	0.461	0.064	0.176
			300	0.015	0.000	0.035	0.025	0.725	0.760	0.225	0.215
			500	0.045	0.005	0.005	0.005	0.547	0.645	0.403	0.345
	2	AUC	100	0.552	0.510	0.667	0.593	0.768	0.587	0.536	0.499
			300	0.500	0.500	0.588	0.587	0.926	0.893	0.882	0.698
			500	0.513	0.500	0.547	0.538	0.968	0.974	0.966	0.951
		MSP	100	0.099	0.271	0.222	0.205	0.557	0.243	0.123	0.281
			300	0.000	0.010	0.015	0.070	0.640	0.730	0.345	0.190

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
			500	0.000	0.000	0.015	0.005	0.512	0.545	0.473	0.450
(ii)	0	AUC	100	0.583	0.508	0.654	0.610	0.774	0.686	0.577	0.535
			300	0.550	0.503	0.580	0.618	0.911	0.871	0.792	0.727
			500	0.668	0.512	0.537	0.553	0.970	0.945	0.931	0.874
		MSP	100	0.040	0.005	0.740	0.840	0.190	0.125	0.030	0.030
			300	0.010	0.000	0.105	0.285	0.800	0.610	0.085	0.105
			500	0.000	0.000	0.000	0.060	0.885	0.755	0.115	0.185
	2	AUC	100	0.547	0.511	0.632	0.596	0.683	0.583	0.546	0.515
			300	0.504	0.500	0.590	0.600	0.851	0.805	0.778	0.675
			500	0.542	0.505	0.536	0.564	0.922	0.869	0.925	0.864
		MSP	100	0.035	0.010	0.835	0.925	0.100	0.055	0.030	0.010
			300	0.005	0.000	0.335	0.630	0.435	0.270	0.225	0.100
			500	0.005	0.000	0.035	0.200	0.620	0.430	0.340	0.370
(iii)	0	AUC	100	0.785	0.542	0.794	0.772	0.862	0.855	0.766	0.694
			300	0.852	0.833	0.840	0.825	0.906	0.900	0.860	0.856
			500	0.881	0.858	0.838	0.844	0.913	0.903	0.877	0.876
		MSP	100	0.365	0.025	0.095	0.105	0.455	0.820	0.085	0.050
			300	0.020	0.000	0.010	0.010	0.955	0.970	0.015	0.020
			500	0.010	0.000	0.025	0.005	0.925	0.980	0.040	0.015
	2	AUC	100	0.754	0.540	0.790	0.799	0.861	0.853	0.708	0.641
			300	0.837	0.782	0.838	0.824	0.894	0.885	0.856	0.830
			500	0.860	0.840	0.844	0.835	0.907	0.899	0.878	0.863
		MSP	100	0.260	0.035	0.140	0.160	0.570	0.765	0.030	0.040
			300	0.005	0.000	0.055	0.035	0.910	0.965	0.030	0.000
			500	0.000	0.000	0.015	0.025	0.940	0.970	0.045	0.005

References

- Berger JO, Wang X, and Shen L (2014). A bayesian approach to subgroup identification. *Journal of biopharmaceutical statistics*, 24(1):110–129. [PubMed: 24392981]
- Breiman L, Friedman J, Stone CJ, and Olshen RA (1984). *Classification and regression trees*. CRC press.
- Dusseldorp E and Van Mechelen I (2014). Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. *Statistics in medicine*, 33(2):219–237. [PubMed: 23922224]
- Foster JC, Taylor JM, and Ruberg SJ (2011). Subgroup identification from randomized clinical trial data. *Statistics in medicine*, 30(24):2867–2880. [PubMed: 21815180]
- Lipkovich I, Dmitrienko A, and R. B. D. Sr. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, 36(1):136–196. [PubMed: 27488683]
- Loh W-Y (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, pages 361–386.

- Loh W-Y (2009). Improving the precision of classification trees. *The Annals of Applied Statistics*, pages 1710–1737.
- Loh W-Y, Fu H, Man M, Champion V, and Yu M (2016). Identification of subgroups with differential treatment effects for longitudinal and multiresponse variables. *Statistics in Medicine*, 35(26):4837–4855. [PubMed: 27346729]
- Loh W-Y, He X, and Man M (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in medicine*, 34(11):1818–1833. [PubMed: 25656439]
- Scheier MF, Helgeson VS, Schulz R, Colvin S, Berga SL, Knapp J, and Gerszten K (2007). Moderators of interventions designed to enhance physical and psychological functioning among younger women with early-stage breast cancer. *Journal of Clinical Oncology*, 25(36):5710–5714. [PubMed: 17998547]
- Schnell PM, Tang Q, Offen WW, and Carlin BP (2016). A bayesian credible subgroups approach to identifying patient subgroups with positive treatment effects. *Biometrics*.
- Sivaganesan S, Laud PW, and Müller P (2011). A bayesian subgroup analysis with a zero-enriched polya urn scheme. *Statistics in medicine*, 30(4):312–323. [PubMed: 21225894]
- Sivaganesan S, Müller P, and Huang B (2017). Subgroup finding via bayesian additive regression trees. *Statistics in Medicine*.
- Su X, Meneses K, McNeas P, and Johnson WO (2011). Interaction trees: exploring the differential effects of an intervention programme for breast cancer survivors. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 60(3):457–474.
- Su X, Tsai C-L, Wang H, Nickerson DM, and Li B (2009). Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(Feb):141–158.
- Su X, Zhou T, Yan X, Fan J, and Yang S (2008). Interaction trees with censored survival data. *The International Journal of Biostatistics*.
- Tian L, Alizadeh AA, Gentles AJ, and Tibshirani R (2014). A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association*, 109(508):1517–1532. [PubMed: 25729117]
- Zhao Y, Zeng D, Rush AJ, and Kosorok MR (2012). Estimating individualized treatment rules using outcome weighted learning. *Journal of the American Statistical Association*, 107(499):1106–1118. [PubMed: 23630406]
- Zhao Y-Q, Zeng D, Laber EB, Song R, Yuan M, and Kosorok MR (2015). Doubly robust learning for estimating individualized treatment with censored data. *Biometrika*, 102(1):151. [PubMed: 25937641]

Figure 1.
Four tree diagrams displaying different structural features.

Figure 2.
Illustrative node assignment matrices for four tree model examples.

Figure 3. Heatmaps from left to right: MSP results using original T-BIC and MSP results using empirical T-BIC.

Figure 4. Heatmaps from left to right: MSP results using original T-BIC and MSP results using empirical T-BIC.

Figure 5. Tree structures produced by different methods for the BCRP data application.

Table 1:

Selected features of IT, QUINT, VT and GUIDE.

	IT	QUINT	VT	GUIDE
Qualitative Interaction Detection	Yes	Yes	Yes	Yes
Quantitative Interaction Detection	Yes	No	Yes	Yes
Variable Selection Bias	Yes	Yes	Yes	No
Handling More Than 2 Treatment Arms	No	No	No	Yes
Computation Speed Rating	Fast	Intermediate	Fast	Fast

Table 2:FSR values and MSP values for models (a)-(d) with $\rho=0$, $\sigma=1$.

			IT		QUINT		VT		GUIDE	
Model	Measure	p n	10	50	10	50	10	50	10	50
(a)	FSR	100	0.064	0.027	0.555	0.742	0.230	0.678	0.013	0.012
		300	0.000	0.000	0.173	0.372	0.550	0.873	0.001	0.002
		500	0.000	0.000	0.026	0.109	0.726	0.643	0.000	0.001
	MSP	100	0.404	0.439	0.170	0.116	0.000	0.000	0.426	0.445
		300	0.354	0.381	0.293	0.239	0.000	0.000	0.354	0.380
		500	0.336	0.345	0.328	0.308	0.000	0.002	0.336	0.345
(b)	FSR	100	0.087	0.029	0.540	0.750	0.196	0.402	0.001	0.001
		300	0.001	0.000	0.161	0.372	0.465	0.545	0.002	0.000
		500	0.000	0.000	0.024	0.099	0.640	0.426	0.001	0.001
	MSP	100	0.395	0.438	0.174	0.112	0.000	0.000	0.431	0.450
		300	0.352	0.381	0.296	0.239	0.000	0.000	0.352	0.380
		500	0.336	0.345	0.328	0.311	0.000	0.000	0.336	0.344
(c)	FSR	100	0.067	0.027	0.563	0.715	0.172	0.427	0.108	0.027
		300	0.000	0.000	0.184	0.379	0.522	0.515	0.351	0.079
		500	0.000	0.000	0.022	0.118	0.615	0.458	0.467	0.170
	MSP	100	0.455	0.440	0.189	0.129	0.002	0.000	0.354	0.431
		500	0.469	0.403	0.382	0.250	0.000	0.000	0.149	0.347
		500	0.471	0.378	0.461	0.334	0.000	0.002	0.068	0.286
(d)	FSR	100	0.132	0.050	0.561	0.721	0.180	0.372	0.033	0.014
		300	0.014	0.003	0.166	0.386	0.397	0.581	0.125	0.028
		500	0.017	0.002	0.025	0.096	0.443	0.436	0.217	0.056
	MSP	100	0.401	0.430	0.172	0.126	0.001	0.000	0.426	0.444
		300	0.394	0.391	0.336	0.241	0.000	0.000	0.269	0.369
		500	0.400	0.355	0.398	0.322	0.001	0.000	0.201	0.322

Table 3:AUC values and MSP values for models (i)-(iii) with $\rho=0$, $\sigma=1$.

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
(i)	0	AUC	100	0.868	0.558	0.739	0.704	0.976	0.977	0.892	0.811
			300	0.992	0.949	0.582	0.569	0.994	0.991	0.993	0.993
			500	0.996	0.994	0.538	0.516	0.996	0.996	0.996	0.996
		MSP	100	0.207	0.045	0.106	0.085	0.410	0.672	0.276	0.199
			300	0.308	0.203	0.018	0.018	0.237	0.309	0.438	0.470
			500	0.303	0.248	0.009	0.004	0.253	0.338	0.434	0.410
	2	AUC	100	0.610	0.506	0.718	0.670	0.896	0.732	0.687	0.590
			300	0.600	0.500	0.595	0.580	0.988	0.984	0.987	0.982
			500	0.810	0.538	0.564	0.543	0.992	0.991	0.992	0.992
		MSP	100	0.099	0.184	0.218	0.262	0.490	0.354	0.193	0.199
			300	0.025	0.000	0.035	0.025	0.376	0.365	0.564	0.610
			500	0.088	0.005	0.025	0.020	0.368	0.368	0.520	0.607
(ii)	0	AUC	100	0.834	0.543	0.717	0.721	0.872	0.863	0.855	0.759
			300	0.968	0.898	0.603	0.556	0.994	0.993	0.992	0.990
			500	0.990	0.971	0.529	0.542	0.996	0.996	0.997	0.995
		MSP	100	0.325	0.045	0.320	0.530	0.125	0.275	0.230	0.150
			300	0.064	0.010	0.050	0.015	0.708	0.738	0.178	0.238
			500	0.176	0.010	0.020	0.020	0.571	0.736	0.234	0.234
	2	AUC	100	0.632	0.531	0.687	0.675	0.781	0.678	0.683	0.574
			300	0.649	0.517	0.619	0.608	0.964	0.889	0.960	0.929
			500	0.832	0.592	0.563	0.529	0.994	0.971	0.990	0.987
		MSP	100	0.105	0.010	0.640	0.825	0.135	0.125	0.120	0.040
			300	0.005	0.010	0.075	0.190	0.570	0.265	0.350	0.535
			500	0.000	0.000	0.015	0.025	0.683	0.505	0.302	0.470
(iii)	0	AUC	100	0.755	0.543	0.786	0.745	0.824	0.809	0.786	0.734
			300	0.883	0.820	0.821	0.812	0.890	0.884	0.882	0.845
			500	0.891	0.855	0.814	0.802	0.900	0.895	0.895	0.876
		MSP	100	0.465	0.060	0.170	0.200	0.220	0.550	0.145	0.190
			300	0.400	0.090	0.015	0.030	0.420	0.820	0.165	0.060
			500	0.295	0.025	0.000	0.015	0.380	0.740	0.325	0.220
	2	AUC	100	0.737	0.526	0.766	0.741	0.798	0.770	0.721	0.627
			300	0.823	0.702	0.835	0.800	0.878	0.858	0.842	0.817
			500	0.843	0.802	0.805	0.796	0.893	0.872	0.875	0.847
		MSP	100	0.375	0.040	0.180	0.250	0.340	0.595	0.105	0.115
			300	0.035	0.000	0.030	0.135	0.820	0.825	0.115	0.040
			500								

				IT		QUINT		VT		GUIDE	
Model	c_0	Measure	np	10	50	10	50	10	50	10	50
			500	0.025	0.005	0.005	0.040	0.685	0.810	0.285	0.145

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 4:

Brief description for all variables from the Breast Cancer Recovery Project. The potential moderators were measured at baseline.

Variable Name	Variable Type	Range		Shorter Name
Potential moderators				
Age	Continuous	29.5	51.3	age
Nationality	Binary	0	1	nationality
Marital status	Binary	0	1	marital
Weight change	Binary	0	1	wcht
Treatment extensiveness index	Continuous	-1.8	2.6	trext
Comorbidities	Continuous	0	13	comorbid
Dispositional optimism	Continuous	1	24	disopt
Unmitigated communion	Continuous	15	42	uncomt
Negative social interaction	Continuous	5	16	negsoct
Outcome				
Improvement in depression Score	Continuous	-23	15	