SDT: A Tree Method for Detecting Patient Subgroups with Personalized Risk Factors

Xiangrui Li, MS[1], Dongxiao Zhu, PhD[1], Ming Dong, PhD[1], Milad Zafar Nezhad, MS[2], Alexander Janke, BS[3], Phillip D. Levy, MD, MPH[4]
[1]Department of Computer Science; [2]Department of Industrial and Systems Engineering; [3]Wayne State University School of Medicine; [4]Department of Emergency Medicine and Cardiovascular Research Institute. Wayne State University, Detroit, MI, USA.

Abstract

Eradicating health disparity is a new focus for precision medicine research. Identifying patient subgroups is an effective approach to customized treatments for maximizing efficiency in precision medicine. Some features may be important risk factors for specific patient subgroups but not necessarily for others, resulting in a potential divergence in treatments designed for a given population. In this paper, we propose a tree-based method, called Subgroup Detection Tree (SDT), to detect patient subgroups with personalized risk factors. SDT differs from conventional CART in the splitting criterion that prioritizes the potential risk factors. Subgroups are automatically formed as leaf nodes in the tree growing procedure. We applied SDT to analyze a clinical hypertension (HTN) dataset, investigating significant risk factors for hypertensive heart disease in African-American patients, and uncovered significant correlations between vitamin D and selected subgroups of patients. Further, SDT is enhanced with ensemble learning to reduce the variance of prediction tasks.

1. Introduction

Due to health disparities, identifying possible subgroups plays an important role in designing treatment schemes and assessing treatment effects for a given individual patient. The subgroups defined by patients' features enable clinicians to explore whether and where heterogeneity of the treatment effect occurs; those features defining subgroups in turn may shed light on the complex relationships between the disease phenotype and patient's risk factors.

In recent years, precision medicine has been brought to great attention. As defined by the National Research Council (NRC)[1], precision medicine is "the tailoring of medical treatment to the individual characteristics of each patient"; methodologically, precision medicine is referred to "the ability to classify individuals into subpopulations that differ in their susceptibility to a particular disease, in the biology and/or prognosis of those diseases they may develop, or in their response to a specific treatment". This implies that accurate identification of patient subgroups and associated risk factors emerge as a promising path to precision medicine. Multi-disciplinary collaboration, such as medical science, statistics and computer science are essential to designing and developing effective subgroup analysis approaches.

Traditional treatment schemes are designed based on the homogenous diagnosis of patients. However, the "one-size-fits-all" approach is not always successful due to ubiquitous differences in treatment effects across and within patient subgroups. One possible reason accounting for that phenomenon is that there exist different risk factors for different patient subgroups, yet the "one-size-fits-all" treatments do not take it into consideration. Therefore, identifying patient subgroups and the specific factors associated with risk and treatment response becomes a major analytical challenge.

Frequently, clinical researchers have found that the selected features of patients are possibly linked to a disease phenotype, yet there is no strong evidence from the conventional whole group analysis to uncover the opaque association. Revealing the subgroup specific linkage could potentially uncover the mechanism of a disease. From the new perspective of subgroup analysis, exploring the opaque association boils down to detecting whether the specific feature(s) are risk factor(s) for a subgroup of patients, but not necessarily for the whole group. Furthermore, if that feature is a key risk factor for a patient subgroup, finding other significant features (risk factors) is also of great importance for clinicians.

There have been many machine learning methods developed to identify, select and prioritize risk factors. Lasso type methods[2,3] are widely used for risk factor selection[4,5] due to the shrinkage effect on feature coefficients; random forest[6] is capable of measuring importance for features and hence can achieve risk factor selection and

prioritization[7]. However, those techniques are built under an assumption that the patient population is homogenous in phenotype that shares the same group of risk factors. Moreover, these techniques are not applicable when a set of hypothesized risk factors exist. Consequently, they are not capable of identifying variabilities in risk factors for patient subgroups.

In this paper, in response to the aforementioned challenges, we propose a novel tree-based method, named as Subgroup Detection Tree (SDT), to detect subgroups with a (pre-given) hypothesized feature possibly being as a risk factor. We developed a novel splitting criterion to grow a SDT. The splitting criterion seeks a split, which leads to the maximal phenotypic variance reduction jointly in the response (phenotype) and hypothesized features (risk factors), and hence links the response and the features. Splits in SDT identify a set of patient features that are closely related both to the hypothesized features and the phenotype. Subgroups are automatically generated as leaf nodes from tree building procedure. Based on the identified subgroups, personalized risk factors can be developed within each subgroup to assist clinician to treat, intervene or prevent disease more effectively.

The rest of this paper is organized as follows. In Section 2, we describe the problem and motivation in a hypertension (HTN) study. In Section 3, we review related works in tree method, including CART and tree-based methods developed for subgroup analysis. In Section 4, we present the subgroup detection tree method in details. In Section 5 we present a case-study of finding personalized risk factors using a hypertension data. In Section 6, we conclude with discussion.

2. Problem Statement

Recent data suggest that lower respective serum 25-OH D (which changes into an active form of the vitamin D) levels may account for a substantial proportion of the greater age-and sex-adjusted cardiovascular risk among African-Americans[8,9]. Within the framework of subgroup analysis, we focus on a study conducted in the Detroit area where the primary interest is to explore the relationships between vitamin D deficiency and cardiovascular disease disparities and to evaluate the efficacy of adjunct vitamin D therapy. In the study, data was collected from a demographic subgroup (African-Americans) that is at high-risk for HTN. Hypertension has been shown to be the single most important contributor to the existing racial differences in life-years lost from cardiovascular disease, explaining close to 50% of the excess risk that exists within the black community[10]. African-Americans experience higher disease prevalence and, especially in males, poorer overall BP (BP) control than their white and Hispanic counterparts. As a result, African-Americans are at increased risk for adverse, pressure related adverse consequences, particularly premature onset of left ventricular (LV) hypertrophy[11]. The left ventricular mass indexed to body surface area (LVMI) on gadolinium-enhanced cardiac magnetic resonance (CMR) was used as a measure of structural heart damage.

Reasons for the glaring disparities in HTN and its pressure-related consequences are myriad with no single sufficiently explanatory variable. However, clear racial differences in vitamin D exist and are largely attributable to the effects of skin pigmentation on conversion of 7-dehydrocholesterol to vitamin D3 by ultraviolet light. Vitamin D deficiency has been linked to incident cardiovascular disease in other, largely white cross-sectional databases but its presence predisposes to the development of HTN in blacks[12].

With those above in consideration, based on the clinical HTN data, our goal is to detect whether there are subgroups among the participating patients showing associations between LVMI and vitamin D. In addition to subgroup identification, we are also interested in finding the associated features through which LVMI is related to vitamin D if there indeed exist patient subgroups showing significant association between LVMI and vitamin D levels.

3. Related Works

3.1 Tree method

The tree-based method (or called recursive partitioning) is a widely used machine learning technique which partitions feature space into mutually exclusive regions. Starting with a single node containing all the samples, the tree is grown by splitting the parent node into two or more child nodes according to some predefined splitting criterion. Within each child node, the partitioning procedure continues until stopping criteria are met.

In general, we may end up with an overly large initial tree, which unavoidably leads to overfitting. A standard routine for addressing this issue is to prune the initial tree. The pruning algorithm seeks a balance between the goodness-of-fit for training samples and model complexity, and generates a sequence of subtrees. The best tree is then selected using validation methods such as cross validation or other statistical approaches.

3.2 Tree methods for prediction task

The classification and regression tree (CART)[13] is one of the most widely used tree methods in statistical learning and data mining. CART with its pruning idea for tree size selection has greatly advanced the application of tree methods. In the growing procedure, CART seeks a splitting pair (X, c) of the parent node (where X is a feature, c is a splitting point associated with) for a binary and univariate partition. For X is continuous or ordinal, all samples with $X \leq c$ are sent to the left child node otherwise sent to the right child node. For X categorical with k levels, c is a level subset, samples with $X \in c$ goes to the left child node and the others goes to the right. The splitting pair (X, c) is obtained by a greedy search among all possible splits that results in the minimal sum of impurity measures of the left and right node.

As the splitting procedure stops, a large initial tree is grown. To avoid overfitting, Breiman et al. proposed a pruning algorithm based on what is called "cost-complexity" criterion[16]. The cost-complexity criterion is essentially a tradeoff between tree size and goodness-of-fit to the training samples. Using the so-called "weakest-link pruning" (which is an elegant implementation using “cost-complexity” criterion), the pruning procedure ends up with a nested sequence of subtrees. The subtree with the best estimated prediction performance (using cross validation or validation dataset) is selected as the fitted model. We refer to [16] for details of pruning. With the selected best tree, the prediction of a new sample falling into some leaf node is made based on the training samples sitting in the same region.

C4.5[14] is another popular tree method for classification. It uses information entropy in the partition criterion and multi-way split in dealing with categorical variables. Different from CART, C4.5 employs a statistical pruning procedure to choose the optimal subtree. Further development of tree methods in classification and regression includes GUIDE[15] and URPCI[16]. These methods seek unbiased splitting feature selection. The idea behind them is to separate the splitting feature selection and splitting point selection from the greedy search, and splitting feature is selected through some statistical procedure such as hypothesis testing.

3.3 Tree methods in subgroup analysis

Tree based methods in subgroup analysis are greatly developed in recent years. One advantage of tree method is that subgroups are objectively formed as leaf nodes in the tree procedure without any prior hypothesis.

Tree-based method for subgroup analysis was first used in the context of censored survival data. Ciampi et al.[17] proposed the “recursive partition and amalgamation” (RECPAM) algorithm. In RECPAM, splits in the tree algorithm are selected based on a greedy choice of a statistic, which measures the heterogeneity of treatment effects between the resultant subgroups (i.e. child nodes). Based on RECPAM’s CART-similar pruning procedure, Negassa et al.[18] further explored RECPAM in its approach to select the best subtree.

Su et al.[19] developed interaction tree (IT) to identity subgroups showing disparities in treatment effects. The splitting criterion in IT is built on a statistical *t*-test, which measures the interaction between treatment and a feature. The split resulting in the most significant test is chosen to grow the tree. To validate an IT, an “interaction-complexity” pruning criterion, which balances the overall interaction of IT and the IT complexity, along with the “weakest-link” strategy, is used. This pruning procedure generates a nested sequence of subtrees and the optimal subtree is chosen via cross-validation or bootstrapping. As subgroups are identified from IT, further analysis can be performed for determining the heterogeneity across all subgroups.

Qualitative interaction tree (QUINT)[20] is a further development in discovering the heterogeneity of effects of two different treatments. The goal of QUINT is to identify the qualitative interaction in addition to the quantitative interaction. QUINT seeks a split maximizing a weighted sum of a measure for difference between two treatment effects and the size of subgroups. In validating the qualitative interaction tree, the pruning strategy and a bias-corrected bootstrap procedure are used to select the optimal subtree. Once the subtree is chosen, qualitative interaction is detected by examining which treatment of the two is better in each subgroup.

Other tree methods developed for subgroup analysis include Loh et al.[21] that extends GUIDE[15] from regression to subgroup analysis by explicitly treating treatment as a predictor in fitting a linear model in each node; virtual twins[22] that combines random forest and CART to form subgroups; model-based recursive partitoning[23] fitting parametric model in tree building; GLMM trees[24], incorporating generalized linear mixed-effect model and tree method for not only subgroup analysis but also estimation of random effects for clusters. We refer to original papers for details on those methods.

3.4 Subgroups detection tree

Previous works in subgroup analysis are mostly developed in comparing effects for different treatments. However, in our problem, we focus on detecting subgroups in which the hypothesized feature(s) are potential risk factor(s). The latter is critical in early stage prevention and intervenes of disease outcome. To achieve this goal, we proposed a tree method called Subgroup Detection Tree (SDT).

Developing new tree based method is very suitable for detecting subgroups where hypothesized features are linked to disease phenotype. First, tree method is an inherently data-driven and nonparametric statistical learning technique. In the biomedical research field, due to the complicated mechanism of disease, the nonparametric nature of tree method may bring advantages over other parametric models. Secondly, tree method is excellent in dealing with interactions among features. This makes tree method very useful in analyzing clinical data, since it is very likely that nonlinear relationships of features are present. Lastly, the automaticity of tree method is suitable in detecting subgroup associated with a potential risk factor as in our goals: each leaf node may define a subgroup; following the decision path along the tree, models can be easily interpreted, possibly providing insights on finding the association between the disease phenotype and risk factors.

4. Methods

To grow a subgroup detection tree (SDT), we follow the conventional tree building procedure as in CART: (1) growing a large initial tree T_0; (2) pruning T_0 to obtain a nested sequence of subtrees (in our case, we used the "cost-complexity" pruning procedure); (3) choosing the optimal subtree by cross-validation or additional samples only assessing prediction performance on the response. The criterion for the best subtree is the mean squared error (MSE) for the response only. Figure 1 illustrates the overall procedure of training a SDT.

4.1 SDT splitting criterion

Suppose that $\{x_1,\cdots,x_p,x_{in}\}$ is the set of features, y is the response. x_{in} is the hypothesized (pre-given) risk factor which is possibly associated with the response y in some way. In some cases, with prior knowledge, x_{in} is not found to be directly associated with y in the entire patient group. But it is possible that x_{in} is an important feature in some subgroups of patients in which x_{in} associates to y through other features. In our case, y is LVMI, x_{in} is vitamin D.

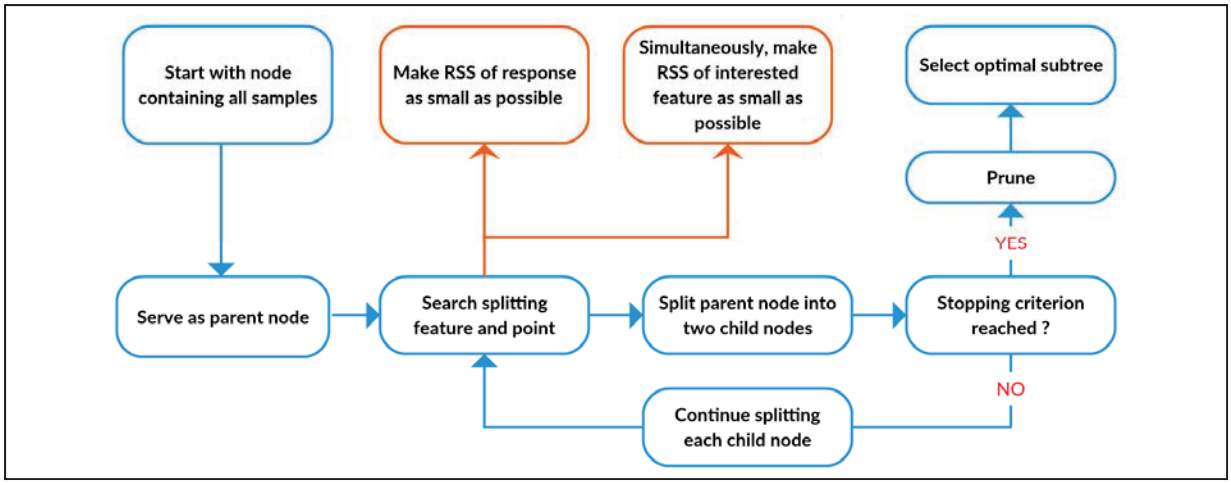

Figure 1. Flowchart of training a SDT. RSS refers to Residual Sum of Square.

Our goal is to (1) detect subgroups in which x_{in} possibly be a risk factor, (2) discover the association features. Within the framework of tree method, those subgroups are defined by the association features. To achieve goal (2), from the perspective of tree method, we will be seeking a splitting pair (X, c) ($X \in \{x_1,\cdots,x_p\}$, x_{in} is not allowed as splitting candidate) that leads to the sum of impurity measures of child nodes for y and x_{in} as small as possible.

We assume that y and x_{in} are continuous. A straightforward choice for node impurity is the residual sum of squares (RSS):

$$I_y(N) = \sum_{i\in N} (y_i - \overline{y}_N)^2, \qquad (1)$$

$$I_{x_{in}}(N) = \sum_{i \in N} (x_{in,i} - \overline{x}_{in,N})^2, \qquad (2)$$

for y and x_{in} respectively, where N is a node, i represents a sample in the node, $\overline{y}_N$ and $\overline{x}_{in,N}$ are the sample mean of y and $\overline{x}_{in}$ respectively.

We will be focusing on binary and univariate split based on the dichotomization of a patient feature. Starting with a feature, say X, and a splitting point c, for X is continuous, whether $X \leq c$ is considered. If a sample answer "yes", it goes to the left child node. Otherwise, it goes to the right child node. For X is a categorical feature, then c is a subset of levels and the splitting rule is samples go to the left child node if $X \in c$ and the right child node if $X \notin c$.

Combining the simultaneousness in minimizing the residual sum of squares of y and x_{in} as the node impurity (1) and (2), we obtain the splitting criterion for seeking a splitting pair (X, c) as follows:

$$\begin{aligned} Q = [\, &\sum_{i \in N_l} (y_i - \overline{y}_l)^2 + \sum_{i \in N_r} (y_i - \overline{y}_r)^2 \,] \qquad (3) \\ &+ w[\, \sum_{i \in N_l} (x_{in,i} - \overline{x}_{in,l})^2 + \sum_{i \in N_r} (x_{in,i} - \overline{x}_{in,r})^2 \,], \end{aligned}$$

where N_l and N_r are the left node and right node respectively, $\overline{y}_l$ and $\overline{y}_r$ are the sample means of y for left and right node respectively, $\overline{x}_{in,l}$ and $\overline{x}_{in,r}$ are similarly the sample means $\overline{x}_{in}$ for left and right node respectively. w denotes the weight for RSS of $\overline{x}_{in}$.

As a side note, $I(N) = I_y(N) + wI_{x_{in}}(N)$ can be viewed as a constrained impurity measure of a node N. If w is set to 0, $I(N)$ is the same with CART and Equation (3) is just equivalent to the splitting criterion for CART in regression. One may treat w as a tuning parameter. In the following analysis of HTN data, we choose $w = \frac{\text{variance of LVMI}}{\text{variance of vitamin D}}$ to put the first and second component of Q into an approximately same scale since the possible maximum for each component is the sample variance.

SDT uses a greedy search for a splitting pair that minimizes the splitting criterion Q. From a single node containing all samples, SDT recursively splits each node until some stopping criterion is reached. At the end, a large initial tree is grown, denoted as T_0.

4.2 Pruning

The initial tree T_0 ending up with the growing process might be very large that probably overfits the training data. Imagine an extreme case in which the SDT allows each node containing a single sample. Then the grown initial tree T_0 can fit the training data perfectly. But it is unlikely to fit future data well. To increase the predictive power of the final tree, we employ the "cost-complexity" pruning idea from CART to SDT. The final tree is then a subtree of T_0.

The cost-complexity function defined in SDT pruning is

$$\begin{aligned} C_\alpha(T) &= \sum_{k=1}^{|T|} [\, \sum_{i \in R_k} (y_i - \overline{y}_k)^2 + w \sum_{i \in R_k} (x_{in,i} - \overline{x}_{in,k})^2 \,] + \alpha|T| \\ &= \sum_{k=1}^{|T|} I(R_k) + \alpha|T| \\ &= C(T) + \alpha|T|, \qquad (4) \end{aligned}$$

where $\{R_1, \cdots, R_{|T|}\}$ is the set of leaf node, $C(T) = \sum_{k=1}^{|T|} I(R_k)$, α is a tuning parameter controlling the tradeoff between the tree size and the goodness of the fitted SDT, T is a subtree of T_0 and $|T|$ is the number of leaf nodes of T. For any α, there is a unique subtree that minimizes C_α. (see below)

A property of node impurity $I(N)$ is that for any node N_p and its child nodes N_{pl} and N_{pr} resulted from any split, the following inequality hold:

$$I(N_p) \geq I(N_{pl}) + I(N_{pr}). \qquad (5)$$

This enables the weakest-link pruning in SDT to adaptively select α as follows.

For $\alpha = 0$, it is obvious from (5) that T_0 minimizes C_α. (It is possible that some subtree T_s satisfies $C_0(T_0) = C_0(T_s)$, then we replace T_0 with T_s as our initial tree). Starting from T_0 and $\alpha = 0$, for any internal node H, denote the subtree rooted at H as T_H. Let $\alpha_1 = \min_H \frac{I(H) - C(T_H)}{|T_H| - 1}$ and H_0 be the internal node corresponding to α_1. Also, denote T_1 as the subtree by pruning off T_{H_0} from T_0. Then we have the following properties: (a) T_1 is the minimal

subtree minimizing $C_{\alpha_1}(T)$. (For the cases that several internal nodes correspond to the same α_1, we prune off all subtrees rooted at those nodes. This ensures that T_1 is unique smallest subtree.) (b) For any α satisfying $\alpha_0 = 0 \leq \alpha < \alpha_1$, T_0 is the minimal subtree minimizing $C_\alpha(T)$.

In other words, we prune off a subtree of T_0 with the smallest per node increase α_1 in $C(T)$. Repeatedly applying this procedure until the trivial tree T_t with one root node only would result in a nested sequence of subtrees $T_0 \supseteq T_1 \supseteq \cdots \supseteq T_t$ and an increasing sequence of α: $\alpha_0 = 0 < \alpha_1 < \alpha_2 < \cdots < \alpha_t < \alpha_{t+1} = \infty$. For any $i \in \{0, 1, \cdots, t\}$, T_i is the unique minimal subtree corresponding to any α with $\alpha_i \leq \alpha < \alpha_{i+1}$.

4.3 Selecting optimal subtree

The final subtree would be selected from the nested sequence of subtrees resulted from pruning. Since SDT aims at discovering the link between the disease and the potential risk factor, the uncovered link is meaningful only when the model performs well on predicting the disease phenotype. Hence, the optimality criterion of selecting the best subtree is chosen as one with the best estimated prediction performance on the response y through some validation methods such as cross-validation or a validation dataset. More specifically, the best subtree in SDT should have the minimal estimated mean squared error (MSE) for the response.

5. Results

In this section, we implement SDT in a clinical dataset to detect whether there are subgroups showing associations between the response and the interested feature. In this data set, response corresponds to LVMI measure and the interested feature refers to vitamin D measure, which is a hypothesized risk factor. Previously, studies have shown that the vitamin D does not highly correlate with LVMI at the whole patient group level (see Table 2). But there may exists subgroups showing high correlation. We show those subgroups can be detected by our new method SDT, but not by the conventional machine learning approaches such as CART.

5.1 Hypertension data information

The clinical data used in our experiment was collected by Detroit Receiving Hospital (DRH) from a group of African-Americans who are at high risk for cardiovascular disease. After data preprocessing and cleaning, there remains 153 samples and 39 features (excluding vitamin D) in the analysis. These features include diabetes history, smoking history, demographic information (gender, ethnicity, education et al.), Cornell product and laboratory results (calcium, chloride, aldosterone, cholesterol, eGFR, parathyroid hormone et al.).

5.2 Experimental result on subgroup detection

To build a SDT, we first used the entire dataset of build a large initial tree T_0, which contains 12 leaf nodes. A node stopped splitting when the size of that node is less than 15. We also set that the minimal size of leaf node is 5. T_0 was then pruned using "cost-complexity" criterion and "weakest-link" procedure back to a trivial tree with a root node only. The pruning procedure resulted in a nested sequence of 9 subtrees. Due to the small sample size (153 samples in the HTN dataset), 10-fold cross-validation was used to estimate the prediction performance only on LVMI. The best subtree was then selected as the one corresponding to the minimal cross-validation MSE of LVMI.

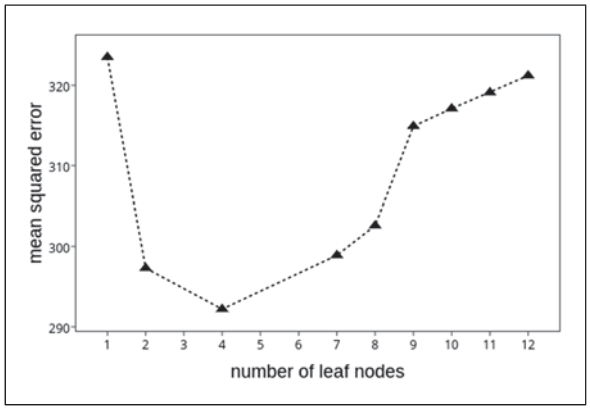

Figure 2. Average of 10-fold cross-validation MSE on LVMI over 100 runs for each subtree.

Since tree method is of high variance, different runs of cross-validation may result in different best subtrees. Therefore, in our analysis, instead of selecting the best subtree from a single run of 10-fold cross-validation, we repeated 10-fold cross validation for 100 times. Each run of 10-fold cross-validation produced a cross-validation Mean Squared Error (MSE) on LVMI for each of 9 subtrees. The best subtree was chosen from 9 subtrees as the one with the minimal average cross-validation MSE on LVMI over those 100 runs. The experiment in this section was performed using customized functions from R package "mvpart".

Figure 2 shows the average MSE of 100 runs of 10-fold cross-validation on LVMI for each subtree (MSE vs. Number of leaf nodes). The minimal average MSE is achieved by the subtree with 4 leaf nodes. Figures 3(a) displays the pruned optimal subtree using SDT.

To examine the performance of SDT on subgroup detection associated with the potential risk factor, we also built a regression tree with CART on the response of LVMI for comparison. R package "tree" was used to build the regression tree. As in selecting the best subtree in SDT, the best subtree of CART is chosen as the one with the minimal average of 10-fold cross-validation MSEs over 100 runs. Figure 3(b) is the resulting best subtree of the regression tree.

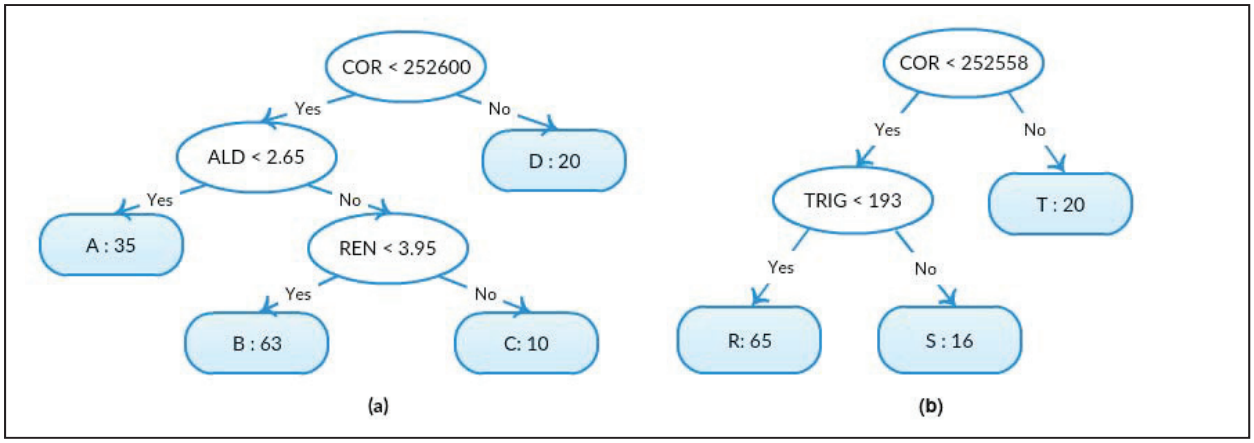

Figure 3. (a) Best subtree for SDT. COR represents Cornell product; ALD is aldosterone; REN refers to renin; (b) best subtree for CART. COR is Cornell Product; TRIG represents triglycerides. For each leaf node denoted as a rectangle, {A, B, C, D} and {R, S, T} are used to label subgroups (leaf nodes) identified by SDT and CART respectively, followed by subgroup size.

For subgroups detected by the optimal subtree, the average LVMI and vitamin D with their standard deviations are calculated. Table 1 summarizes the descriptive statistics for LVMI and vitamin D in SDT and CART. Note that some samples are not sent into subgroups due to their missing values for the selected features.

Table 1. Average of LVMI and vitamin D (along with standard deviation) for subgroups by SDT and CART.

Method	Subgroup	Size	LVMI	Vitamin D
	Entire dataset	153	91.08 (17.93)	11.09(4.01)
SDT	A	35	80.47 (13.31)	9.57 (3.25)
	B	63	94.62 (12.96)	11.49 (3.76)
	C	10	74.98 (9.17)	11.20 (5.07)
	D	20	109.64 (16.61)	10.85 (4.12)
CART	R	65	92.13 (10.85)	10.85 (3.75)
	S	16	99.29 (14.06)	11.44 (4.70)
	T	20	109.64 (16.61)	10.85 (4.12)

To further examine the association between LVMI and vitamin D, association tests using Pearson's correlation coefficient for each detected subgroup were performed: the hypothesis in the tests was chosen as H_0: true correlation $\sigma = 0$ vs. H_a: true correlation $\sigma \neq 0$. Since multiple tests are performed, one may apply Bonferroni-typed adjustment to the resultant *p*-values. The resultant statistics are shown in Table 2 for SDT and CART.

From Table 2, there exists relatively strong negative correlation between LVMI and vitamin D < 10 in Subgroup A and D, indicating that increasing level of vitamin D may decrease LVMI level. Interestingly, in Subgroup C, there is

a positive correlation (0.55) between lower LVMI and vitamin D > 10 suggesting perhaps a threshold effect, possibly mediated by an associated factor that is also modified by vitamin D level such as parathyroid hormone. Figure 4 is the scatter plot (LVMI vs. Vitamin D) for each subgroup, providing a more straightforward illustration for correlation tests.

Table 2. Statistics of correlation tests ($\sigma = 0$ vs. $\sigma \neq 0$) between LVMI and vitamin D for subgroups in SDT and CART. Subgroups of marginal significance are bold-faced. Note that T represents the same subgroup with D.

Method	Subgroup	Correlation	*p*-value
	Entire dataset	-0.12	0.15
SDT	**A**	-0.30	**0.08**
	B	-0.10	0.43
	C	0.55	**0.10**
	D	-0.40	**0.08**
CART	R	-0.07	0.58
	S	0.16	0.54
	T	-0.40	**0.08**

There is a motivation for Pearson's correlation test from the algorithmic perspective. Based on the splitting criterion (3), if the response y and the pre-given feature x_{in} in node N are highly correlated (for example, $\sigma = 0.9$), the split for N could possibly result in large RSS reduction for both the response and the pre-given feature. (Imagine the extreme case that y and x_{in} are linearly related, the optimal split for maximal RSS reduction solely in y is also the optimal split for x_{in}, or *vice versa*.) On the contrary, if the correlation is small, SDT seeks a split that is a compromise in the RSS reduction for y and x_{in}, possibly resulting in much smaller RSS reduction than in the case of high correlation. Since the pruning criterion (4) and (6) collapse subtrees based on the per-node reduction of the sum of RSS of y and that of x_{in}, the pruning procedure tends to keep nodes with high correlation between y and x_{in} (if such nodes were generated in tree growing). This characteristic of the SDT provides a possible explanation that Subgroup A, C and D show relatively high correlations of marginal significance.

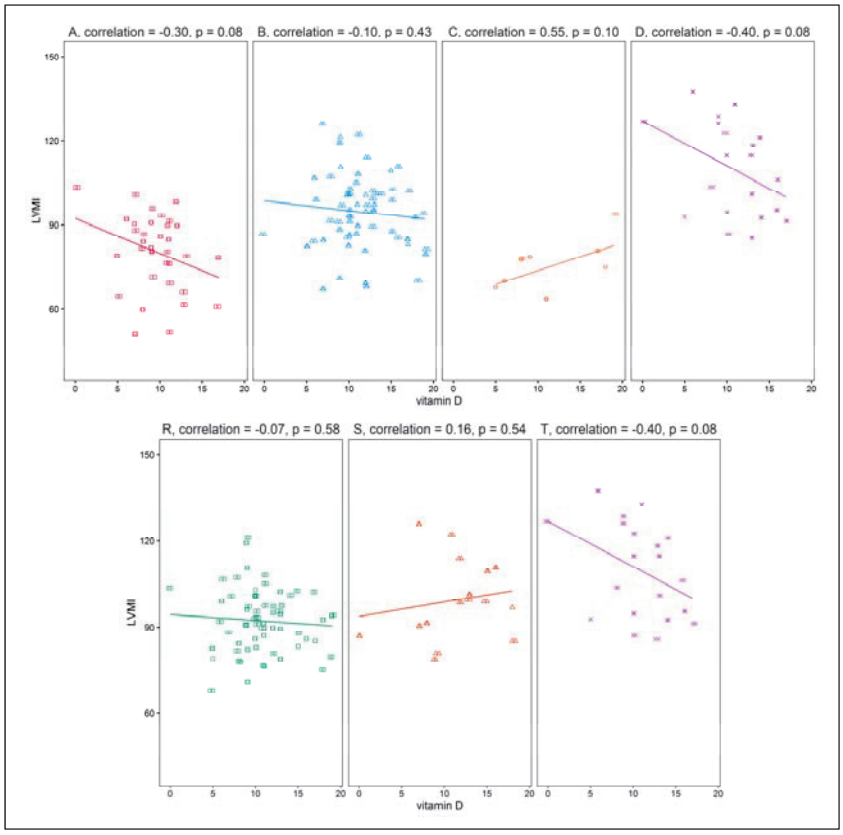

Figure 4. Scatter plots (LVMI vs. Vitamin D) for each subgroup identified by SDT (upper panel) and CART (lower panel).

The first split in the CART selected Cornell product (an

electrocardiographic measure of increased LVMI) as in SDT, leading to a same partition. Since vitamin D component in the splitting criterion of SDT is weighted to a comparable scale with LVMI, Cornell product being selected as the first split indicates that Cornell product is highly correlated with LVMI (an correlation test between LVMI and Cornell product for the entire dataset gives a p-value less than 0.00001). In Subgroup R and S, the correlation between LVMI and vitamin D is small. In contrast, SDT identifies two more subgroups (A and C) showing relatively stronger correlation. This confirms that SDT is more capable of identifying subgroups that are associated with a hypothesized risk factor.

5.3 Prediction performance with bagging SDT

Since SDT can be viewed as a regression tree with a constraint in splitting each node, SDT can be used in the prediction tasks. In general, tree method is known as a supervised model of high variance and data-dependable, so the prediction performance of a single tree may not be desirable. In practice, tree method is often combined with ensemble techniques from machine learning, which could significantly improve the prediction accuracy. Classic ensemble techniques include, to name a few, bagging[25], random forest[6] and boosting[26].

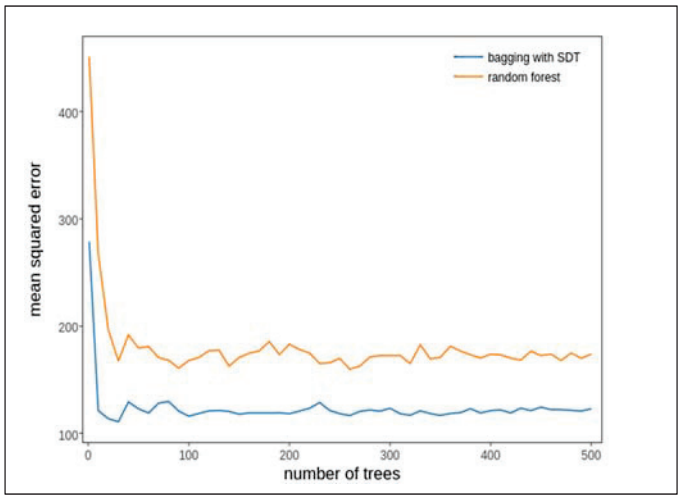

Figure 5. Performance comparison between bagging SDT and random forest.

In this paper, we combined SDT with bagging on the HTN data. In bagging, a bootstrap training set is sampled with replacement from the original dataset; we used this bootstrapping set to train a SDT model without pruning. We repeat this bootstrapping and growing procedure B times (for example, $B = 200$) and obtain a SDT forest. For a future observation, it is fed to every SDT and each SDT B_i produces a fitted value p_i. The final prediction p is then made as the average of those fitted values. That is, $p = \frac{1}{B}\sum_{i=1}^{B} p_i$.

Before testing prediction performance, missing values were imputed with medians for continuous features and the most frequent levels for categorical features. We randomly divided the data into two parts S_1 and S_2. S_1 contains 110 samples used as the training data. The remaining 43 samples in S_2 were used for testing prediction performance. We also ran the data with random forest using R package "randomForest". The comparison of their performances is shown in Figure 5. It is clear that bagging with SDT performs better than random forest on the HTN data.

We randomly divided the data into two parts S_1 and S_2. S_1 contains 110 samples used as the training data. The remaining 43 samples in S_2 were used for testing prediction performance. We also ran the data with random forest using R package "randomForest". The comparison of their performances is shown in Figure 5. It is clear that bagging with SDT outperforms random forest on the HTN data.

6. Discussion

In this paper, we developed a new tree method called SDT for subgroup identification. The SDT tree is grown similarly to CART but in a constrained manner. This constrained approach associates a response and an interested feature by seeking features that are closely related to both. One of the greatest advantages of tree method in linking a subgroup with a specific feature as a risk factor is that each leaf node objectively defines a subgroup without need of prior assumption. Further development is to extend SDT to flexibly accommodate categorical responses, multiple responses, or multiple features in splitting criterion so that SDT can be adapted to solve a wide range of problems in precision medicine.

Particularly in a special case that the constraint in SDT was treated as another response, SDT can be viewed as the regression tree in the multivariate response case. Some works have been done[27], with important difference from SDT in weight handling for each component in splitting criterion and criterion for subtree selection.

Acknowledgement: This paper is based upon work supported by the National Science Foundation under Grant No. 1637312 and 1451316.

References

1. National Research Council (US) Committee. Toward precision medicine: building a knowledge network for biomedical research and a new taxonomy of disease. National Academies Press (US). 2011.
2. Tibshirani R. Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological). 1996 Jan 1:267-88.
3. Zou H, Hastie T. Regularization and variable selection via the elastic net. Journal of the Royal Statistical Society: Series B (Statistical Methodology). 2005 Apr 1;67(2):301-20.
4. Ma S, Song X, Huang J. Supervised group Lasso with applications to microarray data analysis. BMC bioinformatics. 2007 Feb 22;8(1):1.
5. Stuiver MM, Djajadiningrat RS, Graafland NM, Vincent AD, Lucas C, Horenblas S. Early wound complications after inguinal lymphadenectomy in penile cancer: a historical cohort study and risk-factor analysis. European urology. 2013 Sep 30;64(3):486-92.
6. Breiman L. Random forests. Machine learning. 2001 Oct 1;45(1):5-32.
7. Everson TM, Lyons G, Zhang H, et al. DNA methylation loci associated with atopy and high serum IgE: a genome-wide application of recursive Random Forest feature selection. Genome medicine. 2015 Aug 21;7(1):1.
8. Wang L, Song Y, Manson J, et al. Circulating 25-hydroxy-vitamin D and risk of cardiovascular disease a meta-analysis of prospective studies. Circulation: Cardiovascular Quality and Outcomes. 2012 Nov 1;5(6):819-29.
9. Powe CE, Evans MK, Wenger J, et al. Vitamin D–binding protein and vitamin D status of black Americans and white Americans. New England Journal of Medicine. 2013 Nov 21;369(21):1991-2000.
10. Wong MD, Shapiro MF, Boscardin WJ, Ettner SL. Contribution of major diseases to disparities in mortality. New England Journal of Medicine. 2002 Nov 14;347(20):1585-92.
11. Levy P, Ye H, Compton S, et al. Subclinical hypertensive heart disease in black patients with elevated blood pressure in an inner-city emergency department. Annals of emergency medicine. 2012 Oct 31;60(4):467-74.
12. Fiscella K, Franks P. Vitamin D, race, and cardiovascular mortality: findings from a national US sample. The Annals of Family Medicine. 2010 Jan 1;8(1):11-8.
13. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. CRC press; 1984.
14. Quinlan JR. C4. 5: programs for machine learning. Elsevier; 2014 Jun 28.
15. Loh WY. Regression tress with unbiased variable selection and interaction detection. Statistica Sinica. 2002 Apr 1:361-86.
16. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning: A conditional inference framework. Journal of Computational and Graphical statistics. 2006 Sep 1;15(3):651-74.
17. Ciampi A, Negassa A, Lou Z. Tree-structured prediction for censored survival data and the Cox model. Journal of clinical epidemiology. 1995 May 31;48(5):675-89.
18. Negassa A, Ciampi A, Abrahamowicz M, Shapiro S, Boivin JF. Tree-structured subgroup analysis for censored survival data: validation of computationally inexpensive model selection criteria. Statistics and computing. 2005 Jul 1;15(3):231-9.
19. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. Journal of Machine Learning Research. 2009;10(Feb):141-58.
20. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions. Statistics in medicine. 2014 Jan 30;33(2):219-37.
21. Loh WY, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Statistics in medicine. 2015 May 20;34(11):1818-33.
22. Foster JC, Taylor JM, Ruberg SJ. Subgroup identification from randomized clinical trial data. Statistics in medicine. 2011 Oct 30;30(24):2867-80.
23. Seibold H, Zeileis A, Hothorn T. Model-Based Recursive Partitioning for Subgroup Analyses. The international journal of biostatistics. 2016 May 1;12(1):45-63.
24. Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. 2015 Sep.
25. Breiman L. Bagging predictors. Machine learning. 1996 Aug 1;24(2):123-40.
26. Freund Y, Schapire RE. A decision-theoretic generalization of on-line learning and an application to boosting. In European Conference on Computational Learning Theory, Springer Berlin Heidelberg. 1995.
27. De'Ath G. Multivariate regression trees: a new technique for modeling species–environment relationships. Ecology. 2002 Apr 1;83(4):1105-17.