Adv Data Anal Classif
DOI 10.1007/s11634-013-0159-x

REGULAR ARTICLE

A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions

L. L. Doove · E. Dusseldorp ·
K. Van Deun · I. Van Mechelen

Received: 6 March 2013 / Revised: 5 December 2013 / Accepted: 9 December 2013

Abstract In case multiple treatment alternatives are available for some medical problem, the detection of treatment–subgroup interactions (i.e., relative treatment effectiveness varying over subgroups of persons) is of key importance for personalized medicine and the development of optimal treatment assignment strategies. Randomized Clinical Trials (RCT) often go without clear a priori hypotheses on the subgroups involved in treatment–subgroup interactions, and with a large number of pre-treatment characteristics in the data. In such situations, relevant subgroups (defined in terms of pre-treatment characteristics) are to be induced during the actual data analysis. This

Electronic supplementary material The online version of this article (doi:10.1007/s11634-013-0159-x) contains supplementary material, which is available to authorized users.

The research reported in this paper was supported by the Fund for Scientific Research Flanders (G.0546.09N). We would like to thank Xiaogang Su, Ilya Lipkovich and Jared Foster for providing us with the software for Interaction Trees, SIDES and Virtual Twins. In addition, we would like to thank the anonymous reviewers for their valuable comments and suggestions to improve the quality of the paper. For the illustrative applications, we made use of data from a clinical trial conducted as part of the National Drug Abuse Treatment Clinical Trials Network (CTN) sponsored by National Institute on Drug Abuse (NIDA). Specifically, data from CTN-0005 entitled 'MI (Motivational Interviewing) to Improve Treatment Engagement and Outcome in Subjects Seeking Treatment for Substance Abuse' were included. CTN databases and accompanying information are available at www.ctndatashare.org.

L. L. Doove (✉) · E. Dusseldorp · K. Van Deun · I. Van Mechelen
Department of Psychology, Katholieke Universiteit Leuven,
Tiensestraat 102-bus 3713, Leuven,
Belgium
e-mail: lisa.doove@ppw.kuleuven.be

E. Dusseldorp
Netherlands Organisation for Applied Scientific Research TNO,
PO Box 2215, 2301 CE Leiden,
The Netherlands

Published online: 24 December 2013

comes down to a problem of cluster analysis, with the goal of this analysis being to find clusters of persons that are involved in meaningful treatment–person cluster interactions. For such a cluster analysis, five recently proposed methods can be used, all being of a recursive partitioning type. However, these five methods have been developed almost independently, and the relations between them are not yet understood. The present paper closes this gap. It starts by outlining the basic principles behind each method, and by illustrating it with an application on an RCT data set on two treatment strategies for substance abuse problems. Next, it presents a comparison of the methods, hereby focusing on major similarities and differences. The discussion concludes with practical advice for end users with regard to the selection of a suitable method, and with an important challenge for future research in this area.

Keywords Treatment heterogeneity · Recursive partitioning · Subgroup analysis · Treatment–subgroup interaction

Mathematics Subject Classification (2000) 62H30 · 62L99 · 62P10 · 62P15

1 Introduction

For many medical and psychological problems, multiple treatment alternatives are available. A standard research question in such cases pertains to relative treatment effectiveness. A typical setting for the study of such a research question is that of randomized controlled trials (RCTs), in which the persons under study are randomly assigned to different alternative treatment conditions. Beyond some treatment alternative being globally best, relative treatment effectiveness may vary over subgroups of persons that can be characterized in terms of pre-treatment characteristics (i.e., moderators; Kraemer et al. 2002). Formally speaking, relative treatment effectiveness varying over subgroups of persons comes down to a treatment–subgroup interaction. The detection of such interactions may be crucial for personalized medicine and the related development of optimal treatment assignment strategies (Tunis et al. 2010; Dehejia 2005).

Earlier work on the study of treatment–subgroup interactions primarily pertained to situations in which clear a priori hypotheses exist about which subgroups of persons are involved in the interactions, or in which the subgroups can be defined by means of one or a small number of patient characteristics only. Examples include factorial analysis of variance (ANOVA) with a first factor pertaining to treatment methods and a second one to subgroups (Shaffer 1991), and regression analyses with suitable interaction terms being included in the regression model (see, e.g., Dixon and Simon 1991; Hayward et al. 2006).

However, in many RCTs, typically no clear a priori hypotheses exist on the subgroups involved in treatment–subgroups interactions (Bala et al. 2013; Boonacker et al. 2011), and a large number of pre-treatment characteristics are available in the data. In such situations, relevant subgroups (defined in terms of pre-treatment characteristics) are to be induced during the actual data analysis. This comes down to a problem of cluster analysis with the goal of this analysis being to find clusters of persons that are involved in meaningful treatment–person cluster interactions.

For such a cluster analysis, five recently proposed methods can be used, all being of a recursive partitioning type.[1] These are: Model-based recursive partitioning (MOB; Zeileis et al. 2008), Interaction Trees (Su et al. 2008, 2009), Simultaneous Threshold Interaction Modelling Algorithm (STIMA; Dusseldorp et al. 2010), Subgroup Identification based on Differential Effect Search (SIDES; Lipkovich et al. 2011), and Virtual Twins (Foster et al. 2011). Unfortunately, however, these five methods have been developed almost independently, and the relations between them are not yet understood. This is the major challenge we will address in the present paper.

The structure of the remainder of this paper is as follows. In the next section, we will introduce an RCT data set that involves two treatment strategies for persons with a substance abuse problem; we will use this data set throughout this paper for the purpose of an illustrative application. In the third section we will present the five recursive partitioning methods. To this end we will first briefly outline each method, and second we will apply it to the substance abuse data. As our primary interest is in the type of output of each method rather than in a detailed comparison of the results of the different methods on the data of the illustrative example (in terms of specifics such as sizes of the subgroups), we will use in the application the default settings of the tuning parameters of each method under study where possible. In a final section, we will discuss the relations between the five methods, we will give practical advice on how to make a selection between them, and we will list a few challenges for future research.

2 Data

We will reanalyze a data set from the Clinical Trials Network[2] on the evaluation of integrating motivational interviewing techniques into the initial contact and evaluation session of behavior therapies (Carroll et al. 2006). Motivational interviewing has been developed as a treatment strategy to enhance persons' motivation for change. It includes both a motivational interviewing style (e.g., asking open-ended questions, listening reflectively, affirming change-related participant statements and efforts, eliciting self-motivational statements with directive methods, handling resistance without direct confrontation), and motivation-enhancing strategies (e.g., practising empathy, providing choice, removing barriers, providing feedback, and clarifying goals). The data set that we will use pertains to an RCT with participants ($n = 423$) who were seeking treatment for a substance use problem. Following baseline assessment, participants were randomly assigned to one out of two conditions: standard intervention ($n = 214$) and standard intervention in which motivational interviewing techniques were integrated in the intake/orientation sessions ($n = 209$). Due to missing values at follow-up, the available cases for analysis were 178 in the standard and 174 in the motivational interviewing condition. The data comprised 18 pre-treatment character-

[1] A recursive algorithm is defined here as an algorithm involving a stepwise manner of repeating the same procedure.

[2] Clinical Trials Network databases and information are available at www.ctndatashare.org.

istics, including demographical variables (e.g., gender, age, ethnicity) and aspects of substance use (e.g., the primary drug used, and composite scores included in the Addiction Severity Index (ASI; McLellan et al. 1992), the latter being an interview-based measure of the frequency and severity of substance use and related psychosocial problems). As outcome variable, we will focus in our reanalyses on a measure of retention, that is, the number of sessions completed in the 28 days after treatment assignment. The descriptive statistics for all variables are given in Table 1. Previous analyses of these data showed that integrating motivational interviewing techniques in a standard treatment may have a positive effect on retention, especially in the earlier phases of treatment (Carroll et al. 2006). However, the authors of these analyses also hypothesized that treatment effect heterogeneity may be in place, and therefore that 'it is important to understand the types of individuals for whom motivational interviewing is effective …' (Carroll et al. 2006, p. 11). This is exactly the question that will be addressed by the recursive partitioning methods described and illustrated below.

Because of small marginal frequencies for the covariates on primary drug and marital status (benzodiazepines was checked as the primary drug two times only, and widowed was checked as marital status three times only), we merged the categories benzodiazepines and opiates on the primary drug covariate, and the categories widowed and separated for the marital status covariate before applying the recursive partitioning methods to the data. The R-code that has been used to preprocess the data is given in Online Resource 1.

3 Methods

In this section, we will outline the methods of Model-based recursive partitioning, Interaction Trees, STIMA, SIDES, and Virtual Twins. The main ingredients of all methods are: a binary treatment variable T (with in the illustrative application $T = 1$ denoting motivational interviewing and $T = 0$ standard treatment), an outcome variable Y, and a set of pre-treatment characteristics (covariates) X_j $(j = 1, \ldots, J)$, which may be continuous or categorical in nature. All methods are of a recursive partitioning type, which implies that the total group of observations is repeatedly split into child subgroups according to a splitting criterion. The latter differs across the methods. It may be noted that the scope of two of the methods (Model-based recursive partitioning and STIMA) is much broader than that of RCTs and the identification of person clusters involved in treatment–subgroup interactions. For each of these two methods, we will first outline the general framework underlying the method in question; next we will indicate how treatment–subgroup interactions can be addressed within this framework. The R-code for the application of each method to the substance abuse data is given in Online Resource 2, 3, 4, 5 (except for SIDES, the software of which is in the form of an Excel add-in for which the code is not publicly available).

3.1 Model-based recursive partitioning

The idea behind Model-based recursive partitioning is that in many situations, a single global model that fits all observations cannot be found (Zeileis et al. 2008). It might,

Table 1 Descriptive statistics for all variables involved in re-analyses of data from the clinical trials network

Variable	Percent or mean (SD)	
	Standard ($n = 178$)	MI ($n = 174$)
Potential moderators		
Female	39.3	42.5
Ethnicity		
White	70.8	74.7
Other	17.4	14.9
Black, African American, or Negro	8.4	8
Spanish, Hispanic, or Latina	3.4	2.3
Employed	41	33.9
Marital status		
Divorce	18.5	24.7
Living with partner/cohabiting	3.4	3.4
Separated	10.1	9.8
Legally married	18	16.7
Never married	49.4	44.3
Widowed	0.6	1.1
Admission prompted by legal system	55.6	52.3
On probation or parole	38.8	33.9
Any previous drug/alcohol treatment	62.4	60.9
Primary drug used		
Alcohol	51.1	49.4
Cocaine	7.3	5.7
Marijuana	20.2	20.7
Opiates	3.9	5.2
Methamphetamines	16.9	18.4
Benzodiazepines	0.6	0.6
Age	32.8 (10.1)	34.5 (10.4)
Years of education	12.2 (2.1)	12.3 (1.7)
Days of substance use, past 30	10.0 (9.6)	11.6 (10.7)
ASI composite scores		
Medical	0.25 (0.34)	0.25 (0.34)
Employment	0.66 (0.31)	0.68 (0.31)
Alcohol	0.23 (0.25)	0.25 (0.26)
Drug	0.11 (0.11)	0.12 (0.12)
Legal	0.20 (0.22)	0.18 (0.21)
Family	0.16 (0.20)	0.15 (0.19)
Psychological	0.23 (0.23)	0.27 (0.24)
Outcome		
Number of sessions in 28 days after treatment assignment	4.1 (4.1)	5.0 (5.1)

The potential moderators were all measured at baseline (i.e., before receiving standard treatment or motivational interviewing)

SD Standard deviation, *MI* motivational interviewing, *ASI* addiction severity index. Composite scores were calculated according to McGahan et al. (1986)

however, be possible to partition the total group of observations into subgroups on the basis of covariates, so that a well-fitting model can be found for each subgroup. That is, the sample space can be partitioned and local models can be fitted to each partition class.

During the process of Model-based recursive partitioning a tree is grown, every node of which is associated with a parametric model of type $M(O_n, \theta)$, where O_n denotes the observations in node n and θ a vector of parameters, $\theta \in \Theta$. The user has to prespecify the type of the model, (e.g., a logistic regression model representing the regression of a binary outcome variable Y on two predictors), and the group of J covariates X_j ($j = 1, \ldots, J$) to be used in the partitioning.

Given I_n observations $O_{n,i}$ ($i = 1, \ldots, I_n$) in node n, the model $M(O_n, \theta)$ can be fitted by minimizing some objective function $\Psi(O_n, \theta)$, yielding the following vector of parameter estimates:

$$\hat{\theta}_n = \underset{\theta \in \Theta}{\operatorname{argmin}} \sum_{i=1}^{I_n} \Psi(O_{n,i}, \theta). \tag{1}$$

Examples of objective functions include error sums of squares, and minus the log-likelihood. To assess whether it is necessary to split a node, a fluctuation test for parameter instability is performed. That is, it is tested whether the parameter estimates of Model (1) are stable with respect to all partitioning variables $X_1, \ldots, X_J$ against whether splitting the node with respect to one of the partitioning variables X_j may capture instabilities in the parameters and thus improve the fit. The tests used in this step belong to the class of generalized M-fluctuation tests (Zeileis et al. 2008). If there is significant instability with respect to any of the partitioning variables (i.e., p-value for parameter instability falls below some prespecified significance level α), the variable associated with the highest parameter instability is selected. (Note that as in each step of the tree construction a high number of tests has to be performed, Model-based partitioning includes the possibility for a Bonferroni correction against the risk of multiple testing.) The node is then partitioned according to this variable and according to the split point that locally optimizes Ψ in the child nodes. The procedure is repeated in each of the child nodes, again considering parameter instability with respect to all partitioning variables X_j. The algorithm stops when no further instabilities can be found or when a potential split results in a child node that contains less than a prespecified number of observations.

One can use Model-based partitioning to identify subgroups involved in meaningful treatment–subgroup interactions by putting $M(O_n, \theta)$ equal to a regression of treatment outcome (Y) on treatment type (T) while partitioning on the pre-treatment characteristics ($X_1, \ldots, X_J$).

Application We set out the analyses of the substance abuse data by means of Model-based recursive partitioning using all default values for the tuning parameters that control aspects of the fitting algorithm. The latter implies that the tests for parameter instability are Bonferroni corrected, that the significance level α used to determine significant instability is put equal to .05, and that the minimum number of observations in a node is put equal to 20. A linear regression model for number of completed sessions in the first 28 days after treatment assignment explained by treatment is employed and the observations are partitioned making use of all 18 pre-treatment characteristics.

The resulting linear regression-based tree for the substance abuse data is depicted in Fig. 1, using boxplots with fitted regression lines in the leaves. In the fitting process, first a global model for all observations is estimated in the root node. An instability is found with respect to the ASI composite score on psychological problems, which is subsequently used for splitting with an optimal split at score .25. Further parameter instabilities are found in the left child node with respect to the ASI composite score on employment, leading to an optimal split at score .83. No further instabilities with respect to partitioning variables were detected and, hence, the algorithm

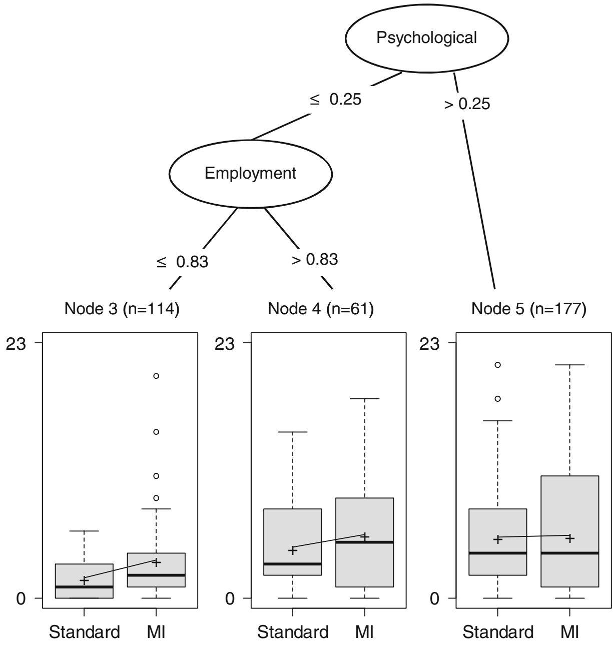

Fig. 1 The linear regression-based tree for the substance abuse data resulting from Model-based recursive partitioning (*MI* motivational interviewing)

Table 2 Summary of the linear-regression-based tree for the substance abuse data

Node	3		4		5	
	($n = 114$)		($n = 61$)		($n = 177$)	
Variable	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	1.81**	0.38	4.55**	0.85	5.48**	0.55
T	1.64*	0.56	1.17	1.18	0.15	0.78

The rows summarize the regressors in the linear regression by means of estimated coefficients and standard errors

T treatment effect of motivational interviewing relative to standard treatment

$^{*}p < 0.01$. $^{**}p < 0.001$

stopped, resulting in three subgroups for whom the effect of treatment on the number of completed sessions differed. Estimates within all leaves of the number of completed sessions for standard treatment (i.e., the intercept) and the effect of motivational interviewing relative to standard treatment (i.e., the slope) are provided in Table 2.

3.2 Interaction Trees

Interaction Trees have been specifically developed to account for heterogeneity of treatment effects (Su et al. 2009). The goal of the method is to account for this heterogeneity by creating subgroups that differ in treatment effect as much as possible. The procedure starts by examining a single split of the data, based on one of the available covariates X_j. Let Z be the binary indicator variable resulting from splitting X_j on a split point c, $Z = 1_{\{X_j \leq c\}}$, in case X_j is continuous in nature, and on a subset c of the values of X_j, $Z = 1_{\{X_j \in c\}}$, in case X_j is categorical in nature. To evaluate the heterogeneity of the treatment effect between the two resulting child nodes, two models are compared: a linear main effects model including treatment variable T and indicator variable Z, and the same main effects model plus the interaction between treatment and the splitting indicator (i.e., $T \times Z$). This is equivalent to a t-test for testing $H_0 : \gamma = 0$ in the following regression model:

$$Y = \beta_0 + \beta_1 T + \delta Z + \gamma T \times Z + E. \tag{2}$$

After an exhaustive search procedure across all partitioning variables and splits, the best variable-split combination is the one that induces the largest effect of the interaction term, that is, the largest difference in treatment effect between the two child nodes. This is quantified in terms of the squared t-statistic for testing $\gamma = 0$ in Model (2), where the t-statistic is squared in order to control for the direction of the comparison. Note that in Model (2) the main effects of pre-treatment characteristics X_j that are not used for splitting are not taken into account. Or, stated differently, in each parent node, Interaction Trees adjust only for main effects of the two predictor variables involved in the treatment–child node interaction.

The splitting process is repeated in each child node until one out of a number of stopping criteria (Su et al. 2009) is met. These stopping criteria include a preset maximum tree depth, and the number of observations in a parent node or in one of the child nodes being under some thresholds that are to be prespecified by the user. The procedure results in a large initial tree, denoted by Q_0.

The best subtree of Q_0 is determined using an interaction-complexity pruning algorithm, based on the pruning algorithm proposed by LeBlanc and Crowley (1993). The main idea is to find the subtree that displays the best balance between amount of interaction and tree complexity. At this point, the amount of interaction of a tree Q (denoted by $G(Q)$) is defined as the sum of the measures of the interaction of all its internal nodes h, denoted by $g(h)$,

$$G(Q) = \sum_{h \in Q} g(h), \tag{3}$$

with

$$g(h) = \left[\frac{\left(\bar{y}_{T=1}^{left} - \bar{y}_{T=0}^{left}\right) - \left(\bar{y}_{T=1}^{right} - \bar{y}_{T=0}^{right}\right)}{\hat{\sigma}\sqrt{\sum_{n=1}^{4}(1/I_n)}} \right]^2, \tag{4}$$

where $n = 1, \ldots, 4$ denotes the four groups resulting from splitting node h (i.e., two child nodes × two treatment groups), I_n is the sample size for group n, and $\hat{\sigma}^2$ a pooled estimate of the variance. Furthermore, the complexity of a tree Q is measured as the number of its internal nodes (denoted by $\#\tilde{Q}$)

To arrive at a subtree of Q_0 with optimal balance between amount of interaction and tree complexity, an iterative procedure is followed. In each step of this procedure a node h will be removed from Q_0, along with the subtree Q_h below it, which is such that for Q_h the ratio $G(Q_h)/\#\tilde{Q}_h$ is minimal. This procedure results in a nested sequence of subtrees, $Q_0 \succeq \ldots \succeq Q_m \succeq \ldots \succeq Q_M$, where $\succeq$ means 'contains as a subtree' and Q_M is the null tree with the root node only. The best-sized tree is selected from this nested subtree sequence by maximizing an interaction-complexity measure defined as:

$$\Upsilon(Q_m) = G(Q_m) - \zeta \cdot \#\tilde{Q}_m, \tag{5}$$

with $\zeta \geq 0$ a parameter to tune the strength of penalizing additional splits; ζ is suggested to be within the range $2 \leq \zeta \leq 4$ by Su et al. (2009).

Application To apply Interaction Trees to the substance abuse data we used the default values for the maximum tree depth (20) and the minimum number of observations per treatment arm in a potential child node (10). As no default has been specified for the minimum number of observations in a parent node, we set this criterion to be 4 times the minimum number of observations per treatment arm in a potential child node (i.e., 40). An initial tree with 8 subgroups is constructed by the algorithm. Along the lines of the example in Su et al. (2009), for the selection of the best subtree, we set the complexity parameter ζ successively equal to 2, 3, 4 and the natural logarithm of the sample size, yielding an unanimously elected best tree size of 8. The tree structure with 8 subgroups is given in Fig. 2. Estimates for Model (2) within each split of the interaction tree are provided in Appendix A.

3.3 STIMA

Within the context of multiple regression models that involve many predictor variables, the goal of STIMA is to automatically search for higher order interaction effects that can be added to a linear regression model with main effects of the predictors under study (Dusseldorp et al. 2010). The higher order interaction effects correspond to the nodes of a regression tree. In STIMA, the multiple regression model with main effects and higher order interactions and the regression tree are estimated simultaneously. A precursor of STIMA, called the regression trunk approach, used a two-step algorithm for this estimation (Dusseldorp and Meulman 2001, 2004).

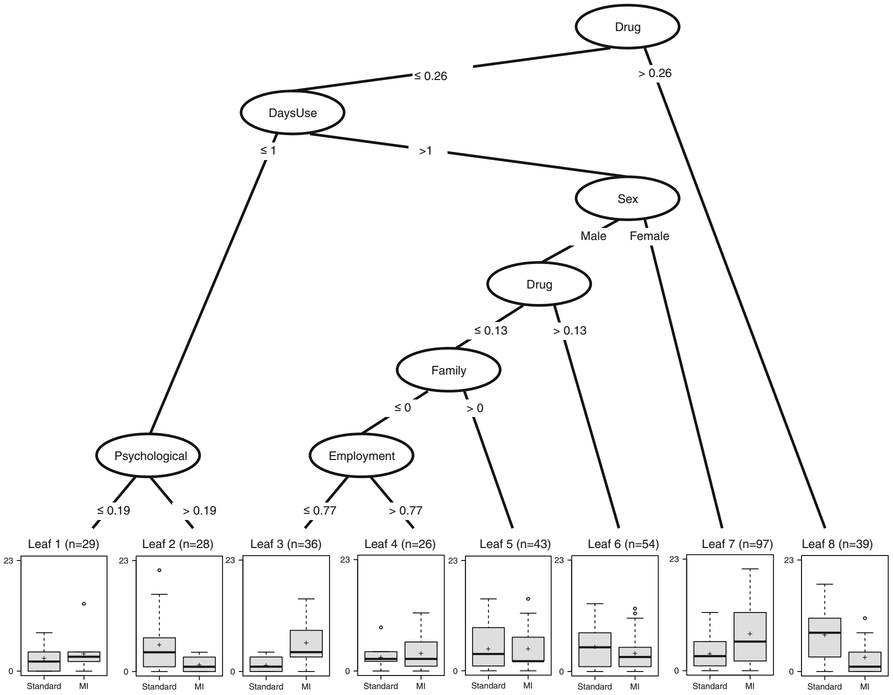

Fig. 2 The best-sized interaction tree for the substance abuse data with in the leaves the *boxplots* for the outcome variable Y in each treatment condition (*MI* motivational interviewing, *DaysUse* days of substance use, past 30)

Given a continuous[3] outcome variable Y and a set of predictor variables X_j ($j = 1, \ldots, J$) that may be continuous or categorical in nature, STIMA constructs a tree based on binary splits of the predictor variables (with each split being such that each of the two resulting child nodes contains more than a to be prespecified number of observations). For this purpose, the algorithm starts in the root node of the tree with the following linear main effects model as the initial reference model:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{J} \beta_j X_j. \tag{6}$$

Note that in Model (6) categorical predictor variables with m values are implicitly replaced by $m - 1$ dummy variables.

To determine the first split, STIMA calculates for each continuous covariate X_{j^*} and for each split point c^* the increase in variance accounted for by the following expanded regression model:

[3] Note that Conversano and Dusseldorp (2010) introduce the use of STIMA in the case of a binary outcome variable Y.

$$\hat{Y} = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + \gamma I(X_{j*} > c^*), \tag{7}$$

in comparison with the variance accounted for by Model (6), where the indicator function $I(X_{j*} > c^*)$ corresponds to a binary split of the data on the basis of predictor X_{j*}. A similar calculation is done for each categorical covariate X_{j^*} for which all values have a marginal frequency ≥ 4 (and each split point c^*), yet with two modifications: (1) The dummy variables corresponding to X_{j^*} are removed from Model (6) and Model (7) to avoid linear dependencies in the latter model; (2) prior to splitting, X_{j^*} is transformed into an ordinal variable by replacing each of its values by the mean residual $Y - \hat{Y}$, with $\hat{Y}$ taken from Model (6), and the mean being calculated across all experimental units that take the value in question. Through an exhaustive search among all covariates and all possible splits, STIMA will ultimately select the covariate—split point combination, say (X_1, c_1), that induces the highest increase in variance accounted for. The root node is split on this combination, and Model (6) is replaced as current reference model by:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + \gamma_1 I(X_1 > c_1), \tag{8}$$

all coefficients of which are re-estimated after the split. Note that, in case of the first split and in case X_1 is categorical, the dummies associated with X_1 are to be removed from the second term at the righthand side of Eq. (8).

In the second step, the exhaustive search is repeated for each of the two child nodes of the root, and that combination (child node, covariate, split point) is chosen that induces the highest increase in variance accounted for. The increase in variance accounted for is calculated on the basis of a further expansion of the current reference model. As an example, for the left child node of the root (defined by $X_1 \leq c_1$), covariate X_{j^*}, and split point c^*, the current reference model is compared with:

$$\hat{Y} = \beta_0 + \sum_{j=1}^{J} \beta_j X_j + \gamma_1 I(X_1 > c_1) + \gamma_2 I(X_1 \leq c_1) I(X_j * > c^*), \tag{9}$$

where the last term now represents a first interaction term. Note that in case X_{j^*} is categorical in nature, it can be considered only as splitting variable provided that in the child node under study all of its values have a marginal frequency ≥ 4; furthermore, in the same case, prior to splitting, X_{j*} is to be transformed into an ordinal variable by replacing each of its values by the mean residual $Y - \hat{Y}$, with $\hat{Y}$ taken from the current reference model.

As a result, the tree is splitted at the optimal (child node, covariate, split point) combination, and the model associated with this combination becomes the new current reference model. The splitting process is repeated until no further suitable splits can be found, or until a to be prespecified maximum number of splits has been reached.[4]

The procedure results in a nested sequence of regression models. The best model among these is determined using V-fold cross-validation. For this purpose, the total data set is partitioned into V subsets S_v. Subsequently, for each subset S_v, STIMA is applied to the data of all objects not belonging to S_v. This leads to a nested sequence of regression models, with for the lth model and the ith object a predicted $\hat{Y}$ value $\hat{y}_{il}^{(v)}$. The cross-validated relative error of the lth regression model then can be estimated as:

$$\mathrm{RE}_l^{cv} = \frac{\sum_{v=1}^{V} \sum_{i \in S_v} \left(y_{i(v)} - \hat{y}_{il}^{(v)} \right)^2}{\sum_{i=1}^{I} (y_i - \bar{y})^2}. \tag{10}$$

Given the lth model, the predicted values from all test sets $\hat{y}_{il}^{(v)}$ can be joined in the vector $(\hat{y}_{il}^{cv}, i = 1, \ldots, I)$. The standard error of the cross-validated relative error estimate of the lth model can then be estimated by:

$$\mathrm{SE}_l^{cv} = \frac{\sqrt{\sum_{i=1}^{I} \left[\left(y_i - \hat{y}_{il}^{cv} \right)^2 - I^{-1} \sum_{i'=1}^{I} \left(y_{i'} - \hat{y}_{i'l}^{cv} \right)^2 \right]^2}}{\sum_{i=1}^{I} (y_i - \bar{y})^2}. \tag{11}$$

Optionally, the cross-validation procedure can be repeated a number of times and the results may be averaged, leading to more stable estimates of RE_l^{cv} and SE_l^{cv}.

The rank number of the regression model with the lowest RE_l^{cv} is denoted by l^*. The rank number of the finally selected model then equals the minimum value of l for which

$$\mathrm{RE}_{l^{**}}^{cv} \leq \mathrm{RE}_{l^*}^{cv} + \rho \mathrm{SE}_{l^*}^{cv}, \tag{12}$$

for some prespecified value of ρ.

To use STIMA for identifying subgroups involved in meaningful treatment–subgroup interactions, the first split in the regression tree is to be made on treatment variable T (this can be forced by the user; Dusseldorp et al. 2010).

Application As just mentioned, in order to detect possible treatment–subgroup interactions, we forced STIMA to use treatment type as the first splitting variable. We used the default value for the minimum number of observations in a node (viz., the square root of the total sample size), while we put the maximum number of splits at 10. We applied the pruning rule with default values for the number of subsets ($V = 10$), the number of times the cross-validation procedure is performed (one), and the value of ρ ($\rho = 0.50$ for $I \geq 300$). This resulted in a regression model without any splits apart

[4] As the same splitting process is repeated in a step-wise manner the STIMA algorithm is recursive according to the definition given in footnote 1. When a more restrictive definition of recursive partitioning is used, whereby what happens in one node from a data-analytic perspective may not depend on information in other nodes, STIMA cannot be called recursive as the splitting procedure in a node of the regression trunk depends on the current reference model and, hence, also on information from other nodes.

from treatment type as the best model; this implies that STIMA found no treatment–subgroup interaction to be present in the substance abuse data.

For illustration purposes only, we will present nevertheless the model with one split in addition to treatment type. The regression coefficients of this model are provided in Table 3, and a plot of the associated regression tree is shown in Fig. 3.

Table 3 The estimated pruned regression trunk model

	Estimate	SE
Main effects		
(Intercept)	1.52	2.34
Female (vs. male)	1.30*	0.54
Black, African American, or Negro (vs. White)	0.62	0.96
Spanish, Hispanic, or Latina (vs. White)	0.44	1.49
Other (vs. White)	1.37*	0.66
Employed (no)	−0.09	0.63
Divorced (vs. separated/widowed)	−0.35	0.90
Living with partner/cohabiting (vs. separated/widowed)	−1.19	1.48
Legally married (vs. separated/widowed)	−0.08	0.94
Never married (vs. separated/widowed)	−0.63	0.84
Admission prompted by legal system (no)	1.88**	0.64
On probation or parole (no)	−0.41	0.63
Any previous drug/alcohol treatment (yes)	0.28	0.53
Alcohol (vs. methamphetamines)	−1.66	0.84
Cocaine (vs. methamphetamines)	−1.06	1.18
Marijuana (vs. methamphetamines)	−1.67*	0.82
Benzodiazephetamines/opiates (vs. methamphetamines)	−0.06	1.23
Age	−0.02	0.03
Years of education	0.15	0.14
Days of substance use, past 30	−0.05	0.03
Medical	−0.54	0.78
Employment	2.40**	1.02
Alcohol	−0.03	1.32
Legal	2.23	1.21
Psychological	1.96	1.23
Drug	−0.27	2.99
Family	−1.48	1.35
Indicator functions		
$I(\text{Treatment} = \text{MI})I(\text{Alcohol} < 0.05)$	−1.56*	0.76
$I(\text{Treatment} = \text{MI})I(\text{Alcohol} \geq 0.05)$	1.90***	0.55

The model is estimated on the total group of observations ($n = 352$). The columns summarize the regression coefficients by means of estimates and standard errors

MI motivational interviewing

$^*p < 0.05$. $^{**}p < 0.01$. $^{***}p < 0.001$

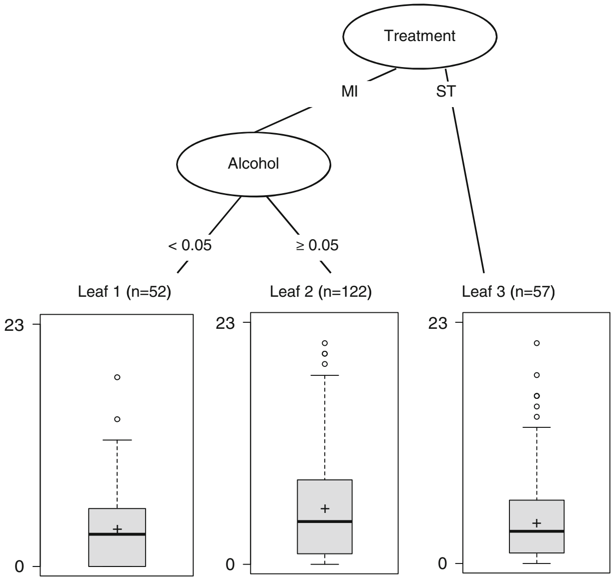

Fig. 3 The regression trunk for the substance abuse data, with enforced interaction effect. The leaves contain *boxplots* for the outcome variable Y (*MI* motivational interviewing, *ST* standard treatment)

3.4 SIDES

SIDES starts by considering one of the two treatments under study as a reference treatment and the other as an alternative treatment. The method then aims at identifying subgroups that are likely to get a high benefit from the alternative treatment in comparison to the reference treatment. This means that SIDES looks for areas in the covariate space where the alternative treatment strongly outperforms the reference treatment, while leaving aside the rest of the space as 'uninteresting' (Lipkovich et al. 2011). The procedure results in a number of (possibly overlapping) subgroups with a relatively higher outcome for the alternative treatment. These subgroups are determined by recursively splitting the data, while selecting multiple splits (M) for each parent node, with M to be prespecified by the user.

For splitting a parent group into two subgroups, only covariates are eligible that are not involved in the definition of that parent node. To determine the M best splits of the parent group under study, one to be prespecified criterion out of a set of three possible splitting criteria is minimized. The three criteria are:

$$w_1 = 2\left[1 - \Phi\left(\frac{|Z^{left} - Z^{right}|}{\sqrt{2}}\right)\right], \tag{13}$$

$$w_2 = 2\min(1 - \Phi(Z^{left}), 1 - \Phi(Z^{right})), \tag{14}$$

$$w_3 = \max(w_1, w_2), \tag{15}$$

where Z^{left} and Z^{right} denote test statistics for a one-sided test of the hypothesis of no differential treatment effect in the left and right child subgroups, and Φ denotes the cumulative distribution function of the standard normal distribution. The first criterion (13) maximizes the difference between the two child nodes in the extent that the alternative treatment outperforms the reference treatment. The second criterion (14) maximizes the extent that the alternative treatment outperforms the reference treatment in at least one of the two child subgroups. The third criterion (15) is a combination of the first and the second. SIDES looks across all combinations of covariates and split points for the M best splits on the selected criterion. Note that several of these splits can be based on the same covariate, yet with other split points.

Within each of the M selected pairs the child subgroup with the smallest p-value for the one-sided test of no differential treatment effect is considered for possible further splitting; if for this test the p-value $< \kappa$, the child subgroup in question is also retained. The value of κ is found using a permutation-based strategy (Lipkovich et al. 2011). This strategy implies that a number of permuted data sets are constructed in which the covariates are decoupled from treatment outcome, by randomly permuting the order of the rows in the covariate part of the data matrix, while leaving treatment type T and outcome Y in place). The SIDES algorithm is then run on each permuted data set for a grid of κ values. Next, for each κ value, the proportion of permuted data sets is calculated for which SIDES retained at least one subgroup (and, hence, erroneously considered at least one covariate relevant for the prediction of differential treatment outcome); we further denote this proportion by $Pr(\kappa)$. The optimal value of κ then is defined as the largest value of κ for which $Pr(\kappa)$ does not exceed a pre-specified nominal level α.

The splitting of SIDES proceeds until at least one of three stopping criteria is met. These criteria pertain to: (1) a maximum for the number of covariates that may be involved in subgroup definitions, (2) a minimum subgroup size, and (3) a minimum improvement in relative level of performance of the alternative treatment in the child subgroup compared to the parent group, with the p-value pertaining to this level being at least a factor of λ smaller in the child subgroup than in the parent group ($0 < \lambda < 1$).

Application We applied SIDES to the substance abuse data introduced in Sect. 2, with motivational interviewing acting as alternative treatment. We used the third splitting criterion (15), and default values for the tuning parameters. The latter implies that the number of best pairs selected at each split M is put equal to 3, that 50 permuted data sets are used and an α-value of .05 for the selection of an optimal value of κ, that the subgroups are described by a maximum of three covariates, that the minimum subgroup size is put equal to 30, and that the relative improvement parameter λ is put equal to .1. With these settings, 20 subgroups with a superior performance of motivational interviewing resulted from the analysis. The three subgroups with the highest value of the test statistic for a one-sided test of the hypothesis of no differential treatment effect are shown in the first row of boxplots in Fig. 4 (the second row provides boxplots for the associated remainder groups of persons). Note that the procedure of selecting multiple splits for each parent node results in subgroups that may be defined by the same covariates with different splits, and that subgroups may overlap. It appears that the variables that are found by SIDES to be most important for predicting a superior performance of motivational interviewing are ‘DaysUse’ and ‘PrimaryDrug’.

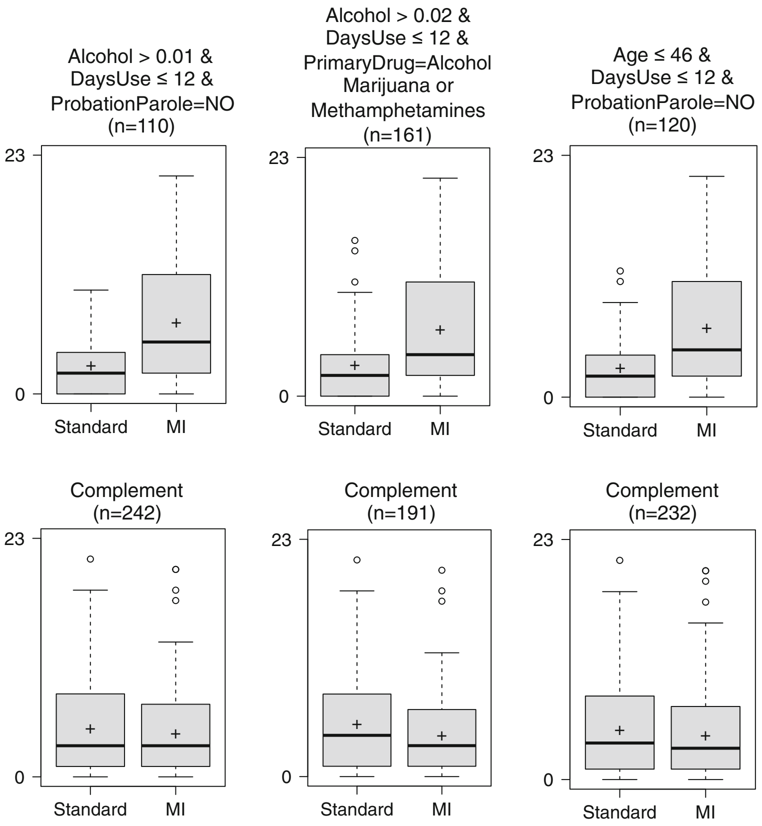

Fig. 4 *Boxplots* of the outcome variable Y for subgroups as determined by SIDES. The *first row* shows boxplots for three subgroups with a high benefit from motivational interviewing in comparison to standard treatment. The *second row* shows *boxplots* for the associated complement group of persons (*MI* motivational interviewing, *DaysUse* days of substance use, past 30, *ProbationParole* on probation or parole, *PrimaryDrug* primary drug used)

3.5 Virtual Twins

Like SIDES, Virtual Twins starts by considering one of the two treatments under study as a reference treatment and the other as an alternative treatment. Virtual Twins then aims at identifying a single subgroup of persons that are likely to get a high benefit from the alternative treatment in comparison to the reference treatment (Foster et al. 2011).

Virtual Twins is based on the concept of counterfactual or potential outcomes (Rubin 1974). The potential outcome of a treatment for a person pertains to the outcome that would be observed if the person in question were subject to that treatment, irrespective of whether this is actually the case or not. As a first step, Virtual Twins estimates for each person the potential outcome values for the alternative treatment and the reference treatment (also referred to as ‘virtual twins’); the difference between these estimates represents for each person an estimate of that person’s individual differential treatment effect. In a second step of Virtual Twins, this estimated individual differential treatment effect is entered as the dependent or criterion variable in a classification or regression tree analysis, together with the available pre-treatment characteristics as predictors. The union of the leaves of this tree for which the average differential treatment effect exceeds some prespecified threshold then constitutes the subgroup that is outputted by the Virtual Twins algorithm.

More specifically, we will further denote the potential outcome for person i under the standard treatment ($T = 0$) and the reference treatment ($T = 1$) by $Y_{i|T=0}$ and $Y_{i|T=1}$ respectively. Estimates of these potential outcomes, $\hat{Y}_{i|T=0}$ and $\hat{Y}_{i|T=1}$, are obtained from a random forest (i.e., a set of regression trees) fitted to the data (Breiman 2001). Virtual Twins constructs random forests of 1000 trees. Each of these trees is based on a bootstrap sample of the persons in the original data as objects, and, as predictor variables of treatment outcome, treatment type (T) and all pre-treatment characteristics (X_j) as well as their interaction with treatment type ($T \times X_j$). During the tree construction, at each node a random subset of J^* predictor variables is considered for splitting that node, with J^* being equal to the total number of predictors divided by 3. Estimates $\hat{Y}_{i|T=k}$ then are derived from the random forest as follows: (1) If k denotes the treatment that person i actually received, $\hat{Y}_{i|T=k}$ is not put equal to the observed Y_i, but to the mean predicted Y-value of the trees that have been constructed on samples that do not contain person i (which is also referred to as an out-of-bag prediction). (2) If k denotes the treatment that person i actually did not receive, $\hat{Y}_{i|T=k}$ is obtained by calculating the mean predicted Y-value after applying the entire random forest to that person's predictor data vector, with the value of T replaced by k. (As a possible alternative procedure, Virtual Twins may also estimate two separate random forests, one for each treatment condition. For the treatment k that person i actually received $Y_{i|T=k}$ can then be estimated by means of the out-of-bag prediction from the random forest of that treatment, whereas for the other treatment it may be estimated on the basis of the other random forest.)

The output of the first step of Virtual Twins is an estimate for each person i of the differential treatment effect for that person, $\hat{Y}_{i|T=1} - \hat{Y}_{i|T=0}$, which we will further denote by $G(i)$. As a second step, Virtual Twins fits a regression tree to the complete data (apart from Y) with $G(i)$ as criterion variable. The leaves of this tree in which the predicted differential treatment effect exceeds a prespecified threshold c then constitute the subgroup A outputted by the Virtual Twins algorithm.

We conclude with two notes. Firstly, the estimate for each person i of the differential treatment effect resulting from the first step of Virtual Twins, $G(i)$, can optionally be dichotomized, making use of a prespecified threshold c^*. In that case, the prediction tree fitted in Step 2 will be a classification tree, and the outputted subgroup A will comprise all leaves in which the majority of the observations takes a value of 1. Secondly, Virtual Twins also includes a module for dealing with binary treatment outcome variables Y.

Application We applied Virtual Twins to the substance abuse data, with Motivational Interviewing acting as alternative treatment. We followed the default option of fitting only a single random forest to the data in Step 1. We further did not dichotomize the estimated differential treatment effect resulting from this step. The regression tree

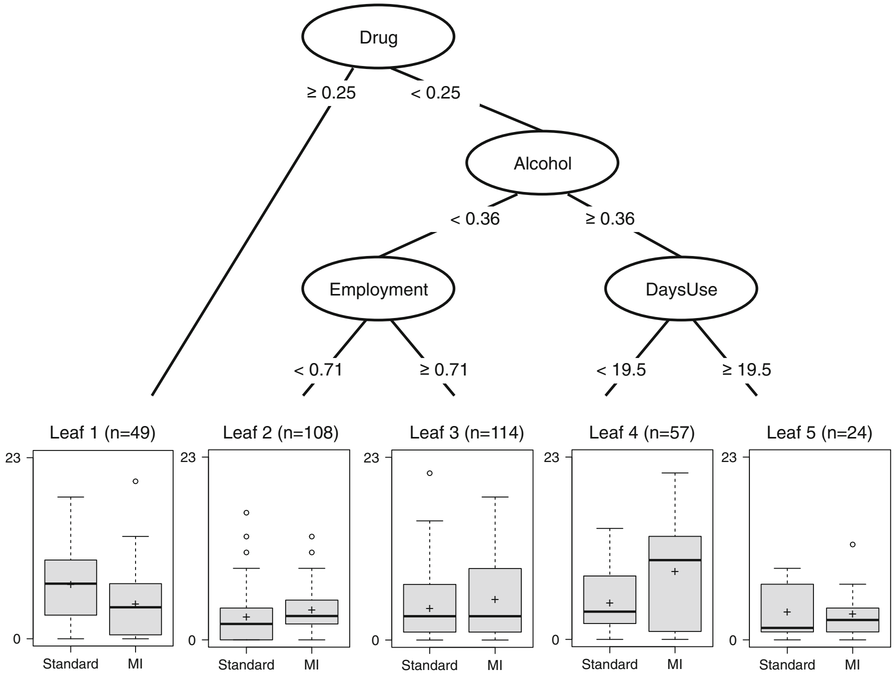

Fig. 5 The regression tree estimated by Virtual Twins. The leaves contain *boxplots* of the outcome variable Y in each treatment condition (*MI* motivational interviewing, *DaysUse* days of substance use, past 30)

resulting from Step 2 is displayed in Fig. 5, using boxplots providing the distribution of Y in the leaves. As threshold for determining which leaves were part of A, we put c equal to one standard error above the overall mean of the estimated differential treatment effect. As this overall mean equalled .94 and its standard error 1.52, this implied that $c = 2.46$. This further implied that A coincides with Leaf 4, which includes 57 persons, or $A = \{i$: Drugs < 0.25 and Alcohol ≥ 0.36 and DaysUse $<$ 19.5$\}$.

4 Discussion

Treatment–subgroup interactions are of major importance for the development of personalized medicine. In common RCT-situations without strong a priori hypotheses and many pre-treatment characteristics, the identification of subgroups involved in such interactions is a major challenge. In the present paper we reviewed five recently proposed methods that can be used to address this challenge.

In view of a comparison on a theoretical/conceptual level, Table 4 shows an overview of the methods with respect to their goals, the type of data they can handle, their underlying model structure, their algorithmic process, and the availability of software. An obvious similarity between the methods is that, in terms of their goal, all of them aim at detecting treatment–subgroup interactions (although the scope of two of the methods is broader than such a detection). A second similarity is that, in terms of the algorithmic process, all methods rely on a recursive partitioning to cluster the persons, with the covariates being split in a binary manner. A major difference between the methods is that the first three aim at representing the full group of persons, whereas the last two focus on the induction of subgroups in which the effect of the alternative treatment is considerably better than the effect of the reference treatment. Moreover, all first three methods (unlike the last two) rely on a regression-type model structure, with treatment outcome as criterion and treatment type as one of the predictors; yet, the way in which they do so differs between them: Model-based recursive partitioning fits local regression models in every node of a tree without incorporating covariate main effects; Interaction Trees fits in each parent node of a tree a local regression model, yet adjusts for the main effect of the covariate used to split that node (and its interaction with treatment type); STIMA constructs a global regression model on the complete data that includes all covariate main effects (and a set of higher-order treatment–covariate interaction terms). The two methods that induce only subgroups with an enhanced relative effect of the alternative treatment differ from one another in terms of the way in which their tree is grown and in terms of the level of complex-

Table 4 Comparison of five recursive partitioning methods in terms of goal, type of data they can handle, underlying model structure, algorithmic process, and availability of software

	MOB	Interaction Trees	STIMA	SIDES	Virtual Twins
Goal					
Identification of subgroups involved in treatment–subgroup interactions	X	X	X	X	X
Specifically designed for study of treatment–subgroup interactions		X		X	X
Represents total group of persons	X	X	X		
Two-sided approach to differences in treatment effect	X	X	X		
Type of data					
Can handle continuous outcome variable	X	X	X	X	X
Can handle categorical outcome variable	X		X	X	X
Underlying model structure					
Regression-based	X	X	X		
Adjustment for covariate main effects		X	X		
Stochastic model	X		X		
Algorithmic process					
Recursive partitioning	X	X	X	X	X
Binary splits on covariates	X	X	X	X	X
Same covariate can be used several times in a branch	X	X	X		X
Pruning		X	X		
Software					
Publicly available	X		X		

MOB Model-based recursive partitioning ity of their output: SIDES dismisses already a large number of subgroups during the tree-growing process, and may yield a rather complex output that comprises a large number of possibly overlapping subgroups; in contrast, Virtual Twins does not dismiss persons during the tree-growing process in which first a large initial tree is grown, after which a single subgroup of persons is selected (that possibly includes several leaves of the tree), which can be considered a simple kind of output.

On an empirical level, we applied the five methods on a data set from the Clinical Trials Network on the evaluation of integrating motivational interview techniques in the initial contact and evaluation session of a therapy for patients with substance abuse problems. Although in this application our primary interest was in illustrating and clarifying the working and type of the output of the respective individual methods, it could nevertheless be useful to look for a synthesis of the different results. Two recurrent elements show up in these. Firstly, especially users without severe drug-related problems seem to benefit from motivational interviewing. However, the nature of the problems that are most relevant as indicators differs across the methods: Model-based recursive partitioning suggests that for motivational interviewing to be more effective than standard treatment, the psychological problems resulting from the drug use should be relatively less severe; Interaction Trees and Virtual Twins indicate that motivational interviewing is relatively more beneficial for persons with less severe drug problems according to the Addiction Severity Index; Interaction Trees also points out that motivational interviewing works relatively better for persons with fewer psychological, family-related and employment problems; from their part, both Virtual Twins and SIDES suggest that fewer days of substance abuse during the month before treatment predicts a higher relative effectiveness of motivational interviewing. Secondly, a reverse situation shows up for alcohol: STIMA, SIDES, and Virtual Twins all three suggest that persons with *more* severe alcohol problems relatively stronger benefit from motivational interviewing. This finding is consistent with the fact that motivational interviewing was initially developed and validated as an intervention for alcohol use disorders (Carroll et al. 2006).

As regards the relevance of the detected treatment–subgroup interactions, we performed a variance component analysis to provide a standardized measure of the effect of the interactions and associated main effects on the number of completed sessions in the first 28 days after treatment assignment as dependent variable. The variance component analysis we set out was based on an analysis of variance model with one factor pertaining to treatment and a second one to the subgroups, using Type I sums of squares. For the two methods that induce subgroups in which the alternative treatment outperforms the reference treatment this implies that the second factor was pertaining to the group of persons with an enhanced relative effect of the alternative treatment and the associated complement group of persons. As statistic for the proportion of variance in the outcome variable accounted for by the interaction and main effects, η^2 is used, which is computed as the ratio of the effect sum of squares to the total sum of squares.[5] The results are given in Table 5, where the proportion of explained variance by SIDES was determined for the treatment–subgroup interaction involving

[5] Given that the total sum of squares is fixed, the η^2 values can be straightforwardly compared across the five methods.

Table 5 Proportions of variance accounted for by the treatment–subgroup interactions adjusted for the main effects of treatment and subgroup, quantified in terms of η^2

Effect	MOB	Interaction Trees	STIMA	SIDES	Virtual Twins
Treatment	0.010	0.010	0.010	0.010	0.010
Subgroup	0.083	0.043	0.001	0.002	0.034
Treatment × subgroup	0.005	0.119	0.017	0.053	0.020

MOB Model-based recursive partitioning

the subgroup with the highest value of the test statistic for a one-sided test of the hypothesis of no differential treatment effect. Two main observations can be made. Firstly, the highest proportions of variance accounted for by the treatment–subgroup interaction were found for the two methods that use differential treatment effect as splitting criterion in the partitioning algorithm, that is, Interaction Trees and SIDES. Interaction Trees further accounts for a considerably higher amount of variance compared to SIDES; this can be explained by the fact that (1) Interaction Trees represents the full group of persons rather than subgroups in which the outcome of the alternative treatment is considerably better and that (2), when using the default setting of the tuning parameters Interaction Trees identifies considerably more subgroups than all other methods for the data at hand.

Next to a theoretical and empirical comparison, one may wonder whether one could give practical advice to end users with regard to making a selection between the different methods. Three questions for the end user seem to be important in this regard: Firstly, the end user needs to decide whether he is interested in representing possible treatment–subgroup interactions for the data at hand as a whole, or rather in identifying one or more subgroups for which the alternative treatment outperforms the reference treatment. Secondly, the end user needs to examine the measurement level of the outcome variable at hand; if this variable is categorical in nature, at present only two of the five methods can handle it. Thirdly, the end user needs to know whether he wants to rely for his analyses on publicly available software; at present only software for Model-based recursive partitioning and STIMA is publicly available.[6]

A final remark pertains to the type of treatment–subgroup interactions addressed by the five methods reviewed in the present paper. The first three methods focus on such interactions in general, whereas the last two methods focus on subgroups in which the alternative treatment outperforms the reference treatment. Yet, none of the methods focuses on so-called qualitative treatment–subgroups interactions, which imply that for some subgroups of persons one treatment is better than the other while for other subgroups the reverse is true. Such interactions are of utmost importance for personalized medicine, and, more in particular, for optimal treatment assignment. Hence, the development of a new, tailor-made method for the identification of sub-

[6] For the analyses reported in this paper we used for Model-based recursive partitioning and STIMA, the `party` (Zeileis et al. 2008) and `stima` (Dusseldorp and Conversano 2013) `R`-code. For the other methods we made use of software kindly made available by the authors of the methods, under the form of `R`-code for Interaction Trees and Virtual Twins, and under the form of an Excel add-in for SIDES.

groups involved in clinically significant qualitative treatment–subgroup interactions looks like an important challenge for future research.

Appendix A

Table 6 provides the estimates for Model 2 within each regression based split of the interaction tree that is shown in Fig. 2.

Table 6 Summary of the regression models at each split of the interaction tree for the substance abuse data

lSplit	Drug		DaysUse		Psychological		Sex		Drug		Family		Employment	
	<= 0.26		<= 1		<= 0.19		= Male		<= 0.13		<= 0		<= 0.77	
	($n = 352$)		($n = 313$)		($n = 57$)		($n = 256$)		($n = 159$)		($n = 105$)		($n = 62$)	
Variable	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE	Estimate	SE
(Intercept)	7.56***	1.07	3.54***	0.40	5.50***	0.94	3.45***	0.68	5.08***	0.77	4.59***	0.78	2.79**	0.81
T	−4.65**	1.46	2.18***	0.56	−4.10*	1.58	4.17***	1.05	−1.36	0.32	−0.02	1.12	0.88	1.19
Z	−3.90***	1.13	0.52	0.85	−2.89*	1.33	0.13	0.84	−2.15*	0.92	−2.71**	1.00	−1.50	1.04
T × Z	6.28***	1.55	−3.71**	1.36	5.03*	2.20	−3.39**	1.15	3.18*	1.29	3.02*	1.46	3.70*	1.57

The columns give for each split the estimated coefficients and standard errors of Model (2)

n size of the node that is split, *DaysUse* days of substance use, past 30, *T* treatment effect of motivational interviewing relative to standard treatment, *Z* splitting indicator, $T \times Z$ interaction between treatment and the splitting indicator

*$p < 0.05$. **$p < 0.01$. ***$p < 0.001$

References

Bala MM, Akl EA, Sun X, Bassler D, Mertz D, Mejza F, Vandvik PO, Malaga G, Johnston BC, Dahm P, Alonso-Coello P, Diaz-Granados N, Srinathan SK, Hassouneh B, Briel M, Busse JW, You JJ, Walter SD, Altman DG, Guyatt GH (2013) Randomized trials published in higher vs. lower impact journals differ indesign, conduct, and analysis. J Clin Epidemiol 66(3):286–295. doi:10.1016/j.jclinepi.2012.10.005. http://www.sciencedirect.com/science/article/pii/S0895435612003174

Boonacker C, Hoes A, van Liere-Visser K, Schilder A, Rovers M (2011) A comparison of subgroup analyses in grant applications and publications. Am J Epidemiol 174(2):219–225

Breiman L (2001) Random forests. Mach Learn 45:5–32

Carroll KM, Ball SA, Nich C, Martino S, Frankforter TL, Farentinos C, Kunkel LE, Mikulich-Gilbertson SK, Morgenstern J, Obert JL, Polcin D, Snead N, Woody GE (2006) Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: a multisite effectiveness study. Drug Alcohol Depend 81(3):301–312. doi:10.1016/j.drugalcdep.2005.08.002

Conversano C, Dusseldorp E (2010) Simultaneous threshold interaction detection in binary classification. In: Palumbo F, Lauro CN, Greenacre MJ (eds) Data analysis and classification, studies in classification, data analysis, and knowledge organization. Springer, Berlin, pp 225–232. doi:10.1007/978-3-642-03739-9_26

Dehejia RH (2005) Program evaluation as a decision problem. J Econ 125(1–2):141–173

Dixon D, Simon R (1991) Bayesian subset analysis. Biometrics 47(3):871–81

Dusseldorp E, Conversano C (2013) stima: simultaneous threshold interaction modeling algorithm. http://CRAN.R-project.org/package=stima, r package version 1.1

Dusseldorp E, Meulman JJ (2001) Prediction in medicine by integrating regression trees into regression analysis with optimal scaling. Methods Inf Med 40:403–409

Dusseldorp E, Meulman JJ (2004) The regression trunk approach to discover treatment covariate interaction. Psychometrika 69(3):355–374

Dusseldorp E, Conversano C, Van Os BJ (2010) Combining an additive and tree-based regression model simultaneously: stima. J Comput Graph Stat 19(3):514–530. doi:10.1198/jcgs.2010.06089

Foster J, Taylor J, Ruberg S (2011) Subgroup identification from randomized clinical trial data. Stat Med 30(24):2867–2880

Hayward RA, Kent DM, Vijan S, Hofer TP (2006) Multivariable risk prediction can greatly enhance the statistical power of clinical trial subgroup analysis. BMC Med Res Methodol 6(18):18. doi:10.1186/1471-2288-6-18

Kraemer H, Wilson G, Fairburn CG, Agras W (2002) Mediators and moderators of treatment effects in randomized clinical trials. Arch Gen Psychiatry 59(10):877–883. doi:10.1001/archpsyc.59.10.877

LeBlanc M, Crowley J (1993) Survival trees by goodness of split. J Am Stat Assoc 88:457–467

Lipkovich I, Dmitrienko A, Denne J, Enas G (2011) Subgroup identification based on differential effect search-a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 30(21):2601–2621

McGahan PL, Griffith JA, Parente R, McLellan AT (1986) Addiction severity index composite scores manual. Treatment Research Institute, Philadelphia

McLellan AT, Al Alterman (1992) A quantitative measure of substance abuse treatments: the treatment services review. J Nerv Mental Dis 180:101–110

R Development Core Team (2013) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. http://www.R-project.org/

Rubin DB (1974) Estimating causal effects of treatment in randomized and nonrandomized studies. J Educ Psychol 66:688–701

Shaffer J (1991) Probability of directional errors with disordinal (qualitative) interaction. Psychometrika 56(1):29–38

Su X, Zhou T, Yan X, Fan J, Yang S (2008) Interaction trees with censored survival data. Int J Biostat 4(1):2

Su X, Tsai CL, Wang H, Nickerson DM, Li B (2009) Subgroup analysis via recursive partitioning. J Mach Learn Res 10:141–158

Tunis SR, Benner J, McClellan M (2010) Comparative effectiveness research: policy context, methods development and research infrastructure. Stat Med 29(19):1963–1976

Zeileis A, Hothorn T, Hornik K (2008) Model-based recursive partitioning. J Comput Graph Stat 17(2):492–514. doi:10.1198/106186008X319331