Aniek Sies* and Iven Van Mechelen

Comparing Four Methods for Estimating Tree-Based Treatment Regimes

DOI 10.1515/ijb-2016-0068

Abstract: When multiple treatment alternatives are available for a certain psychological or medical problem, an important challenge is to find an optimal treatment regime, which specifies for each patient the most effective treatment alternative given his or her pattern of pretreatment characteristics. The focus of this paper is on tree-based treatment regimes, which link an optimal treatment alternative to each leaf of a tree; as such they provide an insightful representation of the decision structure underlying the regime. This paper compares the absolute and relative performance of four methods for estimating regimes of that sort (viz., Interaction Trees, Model-based Recursive Partitioning, an approach developed by Zhang et al. and Qualitative Interaction Trees) in an extensive simulation study. The evaluation criteria were, on the one hand, the expected outcome if the entire population would be subjected to the treatment regime resulting from each method under study and the proportion of clients assigned to the truly best treatment alternative, and, on the other hand, the Type I and Type II error probabilities of each method. The method of Zhang et al. was superior regarding the first two outcome measures and the Type II error probabilities, but performed worst in some conditions of the simulation study regarding Type I error probabilities.

Keywords: treatment regime, subgroup analysis, recursive partitioning, personalized medicine, decision tree

1 Introduction

When multiple treatment alternatives are available for a certain disease, an obvious question is which of these alternatives is most effective. A possible approach to answer such a question, is to set up a comparative treatment evaluation study. Using the data from such a study, a researcher can search for the overall most effective treatment alternative, that is, the alternative with the highest mean outcome across all patients. However, beyond one treatment alternative being on average most effective, it is well-known that the best treatment alternative commonly differs over patients [1, 2]. For example, for one group of patients treatment A may be more effective than treatment B, whereas for another group of patients the reverse may be true.

Due to the presence of such forms of treatment effect heterogeneity, there is a clear need for decision rules that specify for each patient the best treatment he or she should preferably receive, based on patient characteristics that are known prior to treatment. Such decision rules, often referred to as treatment regimes, are a key aspect of personalized medicine, where the focus is on making treatment decisions tailored to the individual patient [3]. Formally speaking, a treatment regime specifies for each patient a treatment alternative given his or her pretreatment characteristics. The optimal treatment regime then is the one leading to the greatest benefit (i.e., the highest expected outcome) if applied to the entire population of patients under study.

Recently, many methods have been proposed to study treatment effect heterogeneity, and can be used to estimate optimal treatment regimes (for a comprehensive review, see Lipkovich et al. [4]; for a focused review and a comparison of several methods on an empirical data set see Doove et al. [5]). These methods differ in several respects, including the type of data on which the estimation is based, the nature of the decisions that are to be made, the number of decision alternatives, and the class of treatment regimes under study.

***Corresponding author: Aniek Sies,** Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, box 3713, 3000 Leuven, Belgium e-mail: aniek.sies@kuleuven.be

Iven Van Mechelen, Faculty of Psychology and Educational Sciences, KU Leuven, Tiensestraat 102, box 3713, 3000 Leuven, Belgium

Regarding the data, these may come from Randomized Clinical Trials (RCTs) or observational studies with patients being assigned to one of the treatment alternatives. With regard to the nature of the decisions, these may pertain to a single, static choice between the treatment alternatives at the start of the treatment as a whole, or to a sequence of dynamic choices during the course of the treatment. In the latter, dynamic case, decisions are to be taken during the treatment, taking into account all accrued information at each decision point [6–8]. As to the decision alternatives, these may be two or strictly greater than two in number. Regarding the class of treatment regimes, one should note that a treatment regime can take different forms. For example, one class includes regimes that take the form of a hyperplane, represented by a weighted linear combination of predictor variables, such as $I(\beta_3 + \beta_4 X1 + \beta_5 X2 > 0)$ [9–12]. Another class comprises regimes based on rectangular regions that may be represented by a decision tree, in which binary splits are made based on dichotomized pretreatment characteristics, and in which the subgroups underlying a regime correspond to the leaves of the tree [13–20], by decision lists [21], or by subsets of the covariate space [22–25]. Still other classes of regimes may rely on nonlinear or generalized linear classifiers [14, 26, 27].

In this paper, we focus on methods for estimating treatment regimes from data of RCTs with regard to a static choice between two treatment alternatives, with the regimes in question being of a tree-based type. Tree-based treatment regimes have the advantage of providing a straightforward and most insightful representation of the decision structure underlying the regimes under study.

During the past few years, quite a few methods have been developed that may be used to estimate tree-based treatment regimes, namely Virtual Twins (VT: Foster et al., 2011 [18]), Model-based Recursive Partitioning (MOB: Zeileis et al., 2008 [17]), Interaction Trees (IT: Su et al., 2009 [16]), Qualitative Interaction Trees (QUINT: Dusseldorp and Van Mechelen, 2014 [13]), Outcome Weighted Learning (OWL: Zhao et al., 2012 [26]), a classification-based approach to estimating optimal treatment regimes: Zhang et al., 2012a [14], Minimum Impurity Decision Assignments Decision Trees (MIDAS: Laber and Zhao, 2015 [15]), and Generalized Unbiased Interaction Detection and Estimation, interaction (Gi: Loh et al, 1015 [19]). These methods have different objectives and optimization criteria, yet, the output of each of them may be transformed into a tree-based treatment regime. In particular, regarding the objective, the method of Zhang et al. [14], OWL and MIDAS have been specifically developed to directly estimate tree-based treatment regimes, unlike Gi, QUINT, IT, MOB and VT; yet, a tree-based treatment regime can be deduced from the output of the latter methods by assigning the treatment with the highest estimated mean outcome to each leaf of the tree in the model output. Regarding the optimization criterion, MOB models the outcome via a recursive partitioning procedure that splits each leaf of the current tree on the basis of the variable and the split point that imply maximal parameter instability; from their part, IT, QUINT and Gi aim at maximizing treatment-subgroup interactions, whereas the classification approach of Zhang and MIDAS maximize a criterion of expected (potential) outcome.

Given all these methods that can be used to estimate tree-based treatment regimes, one may obviously wonder which of these methods results in the best estimated regimes. Unfortunately, however, up to now only partial information is available on the absolute and relative performance of the methods in question. The present paper will address this issue by means of a simulation study in which we will compare the regimes that the methods return. The methods that will be included in this simulation study are methods that can be used to estimate tree-based treatment regimes and for which source code is publicly available. These are MOB, IT, QUINT and the method of Zhang et al. [14] [1]; we will evaluate the performance of these methods with respect to a number of evaluation aspects. Obvious evaluation aspects in this regard firstly pertain to the performance of the resulting regimes. This can be measured in terms of the average outcome of each regime, that is, the expected outcome if all members of the population under study would be assigned to the treatment alternative implied by that regime. Alternatively, one may also look at the proportion of patients that, based on that regime, would receive the truly best treatment alternative from the viewpoint of the underlying true population model. Secondly, one may observe that in the study of treatment effect heterogeneity and

1 Although VT and OWL satisfy these requirements as well, we will not include them in our simulation study because they can be viewed as specific yet less efficient cases of the framework of Zhang et al. [14], see Zhang et al. [14, p. 109] and Lipkovich et al. [4])

subgroup analyses several critics pointed at the ubiquitous presence of Type I and Type II inferential errors [28–32], which some critics even labelled as a "clinicostatistical tragedy" [33]. All inferential errors in question pertain to testing the null-hypothesis of absence of a treatment-subgroup interaction. Within the context of evaluating estimated treatment regimes, two related kinds of error can be distinguished. Here, the null hypothesis pertains to the hypothesis that everyone should receive the same (marginally best or standard) treatment alternative. A first kind of error is committed when H0 is wrongly 'rejected', meaning that a non-trivial tree (i.e., a tree with more than a root node only) for which the optimal treatment alternative differs between leaves, is returned whereas according to the true underlying population model, all patients should receive the same treatment alternative. A second kind of error is committed when H0 is wrongly retained, meaning that a trivial tree (i.e., a tree with a root node only) or a non-trivial tree in which all leaves are assigned to the same treatment alternative, is returned, whereas according to the true underlying population model, patients should be assigned to different treatments. Although strictly speaking incorrect (since no hypothesis is being tested), we will still refer to these errors as Type I and Type II errors. Taking all this into account, we will include Type I and Type II error rates in our simulation study as a second group of evaluation characteristics.

In the absolute and relative evaluation of the four methods under study we want to take into account characteristics of the data and the data-generating mechanism, with the result of the evaluation possibly depending on these characteristics. At this point, in addition to obvious candidates such as sample size and effect size, we want to experimentally manipulate in our simulation study the true treatment regime underlying the generated data. In particular, to understand the performance of the methods under study both when the class of regimes they are based on matches the truth and when it does not, we will consider true tree-based and true non tree-based treatment regimes; additionally we will distinguish between simple- and complex versions of the regimes in question. Furthermore, to investigate the Type I error rate, we will also include a quantitative treatment-subgroup interaction scenario, in which the data are generated on the basis of a tree, the leaves of which differ only in terms of the magnitude and not the sign of the difference in outcome between the treatment alternatives; this implies that, in terms of an optimal treatment regime, everyone should be assigned to the same (marginally best) treatment alternative (which comes down to the treatment regime of the null hypothesis).

The structure of the remainder of this paper is as follows. In Section 2, the four selected methods for estimating tree-based treatment regimes will be briefly outlined. In Section 3 we will describe the method of the simulation study and in Section 4 we will report its results. In Section 5 we will present an illustrative application and finally in Section 6 we will present some concluding remarks.

2 Methods for estimating tree-based treatment regimes

In this section, we will introduce a conceptual framework for the estimation of optimal tree-based treatment regimes. Subsequently, we will briefly outline the four estimation methods focused on in the present paper. (For a more detailed explanation we refer to the articles in which the methods were originally proposed [17, 16, 14, 13]).

2.1 Conceptual framework

The type of data we focus on in this paper, are RCT data where n patients receive at random one out of two available treatment alternatives, denoted by $A = 0$ and $A = 1$. The data include pretreatment characteristics, denoted by $\mathbf{X} = (X_1, ..., X_j, ..., X_J)$. The observed outcome of interest is Y, where we assume that larger values are better. This results in the observed data $(\mathbf{X}_i, A_i, Y_i)$, $i = 1, ..., n$, which are assumed to be independent and identically distributed (IID) across i.

From these data a treatment regime g must be estimated. A treatment regime is a mapping from the value set of $\mathbf{X}$ to the value set of A, meaning a client with a pattern of pretreatment characteristics $\mathbf{X}$ will

receive $A = 0$ if $g(\mathbf{X}) = 0$ and $A = 1$ if $g(\mathbf{X}) = 1$. To formalize the concept of an optimal treatment regime, we need the notion of potential outcomes Y^a, $a = 0, 1$, adopted from the Neyman-Rubin causal model [34, 35]. Potential outcomes Y^0 and Y^1 denote the outcomes for a patient that would have been observed, had the patient received $A = 0$ or $A = 1$, regardless of the actually received treatment. The potential outcome for a patient if he/she would receive treatment according to treatment regime g, would then be:

$$Y^{g(\mathbf{X})} = Y^1 g(\boldsymbol{X}) + Y^0\{1 - g(\boldsymbol{X})\}. \tag{1}$$

The optimal treatment regime within a given class of treatment regimes, G, then is the regime that maximizes the expected potential outcome, when used to assign treatment to all patients in the population of interest:

$$g^{opt}(\mathbf{X}) = \arg\max_{g \in G} E\{Y^{g(\mathbf{X})}\}. \tag{2}$$

In general, several classes of treatment regimes, G, may be distinguished. The class of treatment regimes we are focusing on in this paper is the class of tree-based treatment regimes, G_T, with binary splits based on dichotomized pretreatment characteristics. The splits result in a set of nodes, including a root node, internal nodes, and terminal nodes or leaves. Decisions with regard to treatment assignment are based on leaf membership. Figure 1 shows a hypothetical example of such a tree-based treatment regime, based on an RCT with 422 patients, in which two treatments are being compared. Splits are based on pretreatment characteristics X_5 and X_2. According to the regime, patients with a score lower than 0 on X_5, or with a score higher than 0 on X_5 *and* a score higher than 22 on X_2 should receive Treatment 1. For patients with a score higher than 0 on X_5 *and* a score lower than 22 on X_2, Treatment 0 is the optimal treatment. Below, we will explain how each method under study constructs such a regime.

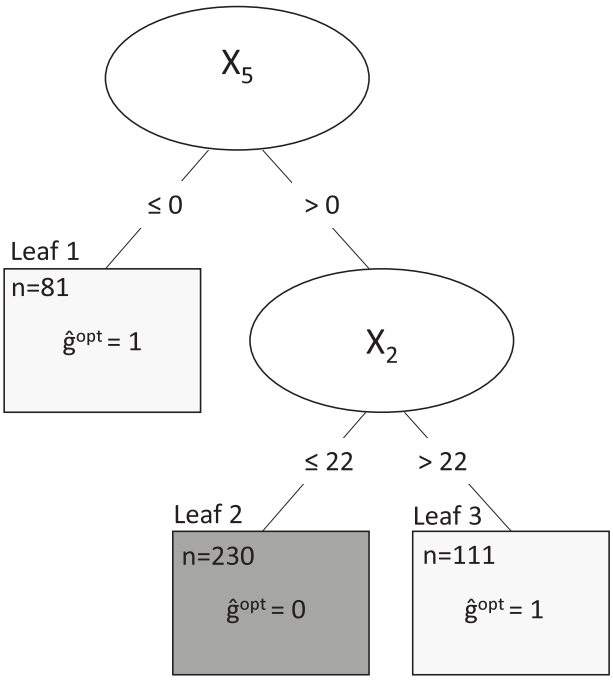


Figure 1: Hypothetical example of a tree-based treatment regime, based on an RCT with 422 patients. Binary splits are made on pretreatment characteristics X_5 and X_2. For patients in Leaf 1 and Leaf 3, the optimal treatment is Treatment 1, whereas for patients in Leaf 2 the optimal treatment is Treatment 0.

2.2 Model-based recursive partitioning

MOB is built on the assumption that in many situations it is unlikely that a single global model fits all observations well. It may then be possible, however, to recursively partition the observations in subgroups based on covariates such that a well-fitting model can be found locally for each subgroup. This general framework can be used to study treatment effect heterogeneity. In this case, the fitted model in each subgroup will be a regression of outcome on treatment, where the intercept and slope are allowed to vary across leaves [36]. A treatment regime can be deduced from the output by examining the sign of the slope; when it is positive, the leaf is assigned to Treatment 1, when it is negative it is assigned to Treatment 0.

In the first step of the recursive partitioning procedure, a regression model $E(Y|X, A) = \beta_0 + \beta_1 \times A$ is fitted to all observations in the current node. This model is referred to as $M(Y, \boldsymbol{\theta})$, where Y denotes all observations, and $\boldsymbol{\theta}$ a vector of parameters, $\boldsymbol{\theta} \in \Theta$. Given n observations, the model, $M(Y, \boldsymbol{\theta})$ can be fitted by minimizing the objective function $\Psi(Y, \boldsymbol{\theta})$ (e.g., the error sum of squares in ordinary least squares or minus the log-likelihood in maximum likelihood estimation), yielding the parameter estimate $\hat{\boldsymbol{\theta}}$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}\in\Theta} \sum_{i=1}^{n} \Psi(Y_i, \boldsymbol{\theta}). \tag{3}$$

To determine whether splitting the node is necessary, it must be tested whether the parameters of the fitted regression model are stable over each ordering implied by the pretreatment characteristics $X_1, \ldots, X_J$, versus whether splitting the observations with respect to one of the characteristics might capture instabilities in the parameters and thus improve the fit. This is tested using a fluctuation test for parameter instability (generalized M-fluctuation test [37]). If there is significant instability with respect to any of the pretreatment characteristics, the node is split according to the characteristic that is associated with the highest parameter instability, and the split point that locally optimizes Ψ is chosen.

This procedure is repeated in each child node, and stops when no further parameter instabilities can be found or when the child node contains less than a pre-specified number of observations. The final tree is automatically selected as the best tree; there is no pruning procedure to correct for over-fitting, because in the parameter instability tests, multiplicity adjusted p-values are used to prevent this. Note that MOB may model main effects of pretreatment characteristics if instability primarily pertains to the intercept, and quantitative treatment-subgroup interactions (implying that the magnitude of the treatment effect differs between subgroups, whereas the optimal treatment alternative does not [38]) if instability is detected in the treatment effect, with the effect being in the same direction in all leaves.

2.3 Interaction trees

IT has been developed to capture treatment-subgroup interactions by recursively splitting the group of patients based on pretreatment characteristics, such that in each split the treatment-split interaction is maximized. This method was not specifically developed for estimating optimal treatment regimes, yet a treatment regime can be deduced from the final tree by assigning each leaf to the treatment with the highest mean outcome in that leaf.

The recursive splitting procedure starts by considering a single possible split, s, of the data based on one of the pretreatment characteristics, X_j. The interaction between split s and treatment is estimated by:

$$h(s) = \frac{(\bar{y}_L^1 - \bar{y}_L^0) - (\bar{y}_R^1 - \bar{y}_R^0)}{\hat{\sigma}\sqrt{1/n_1 + 1/n_2 + 1/n_3 + 1/n_4}}, \tag{4}$$

where $\bar{y}_L^a$ and $\bar{y}_R^a$ are the mean outcomes under treatment a (where $a = 0, 1$, in the left and right child node respectively), $n_1, \ldots, n_4$ are the sample sizes in both child nodes for Treatment 0 and Treatment 1, and

$\hat{\sigma}^2 = \sum_{i=1}^{4} \frac{n_i - 1}{\sum_{j=1}^{4} n_j - 1} s_i^2$ is a pooled estimator of the within-node variance. The best split $s^\star$ is the one yielding the maximum value of $[h(s)]^2$ among all possible splits s. Note that if $(\bar{y}_L^1 - \bar{y}_L^0)$ and $(\bar{y}_R^1 - \bar{y}_R^0)$ are in the same direction, IT is picking up a quantitative treatment-subgroup interaction. Subsequently, the data in node k are split according to $s^\star$, and the procedure is repeated for the resulting child nodes. Recursively doing so results in a large initial tree denoted by T_0.

The splitting procedure ends if either all covariates in a node have the same value, if the total number of observations in the node is less than some predefined minimum node size, if the depth of the node is greater than some predefined maximum tree depth, or if for all permissible splits the minimum of (n_1, n_2, n_3, n_4) is below some predefined threshold. The large initial tree T_0 will be pruned based on an interaction-complexity procedure (which is analogous the the split-complexity pruning algorithm of Leblanc and Crowley [39]). Following this procedure, a nested sequence of trees is created, and the final tree will be the tree from this sequence with the maximum value of an interaction-complexity measure. An honest estimate of this measure can be obtained in two ways, namely by calculating it from an independent test set (an independent subset of the data), or by using a bootstrap procedure in the case of a moderate or low sample size.

2.4 Optimal treatment regime identification by Zhang et al.

The method of Zhang et al. [14] has been developed to directly estimate optimal treatment regimes from observational studies or RCTs. For each patient under study a contrast is estimated,

$$C(\boldsymbol{X}_i) = \mu(1, \boldsymbol{X}_i) - \mu(0, \boldsymbol{X}_i), \tag{5}$$

where $\mu(a, \mathbf{X}) = E(Y|A = a, \mathbf{X})$. This contrast denotes the difference between the expected outcome under Treatment 1 versus Treatment 0, given the patient's pretreatment characteristics. The sign of this contrast indicates the estimated optimal treatment, whereas its absolute value indicates the cost if this patient would be misclassified. Subsequently, the contrasts can be used as input for supervised classification methods to estimate an optimal treatment regime. Note that the choice for the estimator of the contrast function as well as the choice of the supervised classification method are up to the user, making this approach very flexible.

The contrast function $C(\boldsymbol{X}_i)$ can be estimated in various ways. Zhang et al. [14] introduce an augmented inverse probability weighted estimator (AIPWE [40]) as a promising possibility:

$$\hat{C}_{AIPWE}(\boldsymbol{X}_i) = \frac{A_i}{\pi(\boldsymbol{X}_i, \hat{\gamma})} Y_i - \frac{1 - A_i}{1 - \pi(\boldsymbol{X}_i, \hat{\gamma})} Y_i - \frac{A_i - \pi(\boldsymbol{X}_i, \hat{\gamma})}{\pi(\boldsymbol{X}_i, \hat{\gamma})} \mu(1, \boldsymbol{X}_i, \hat{\boldsymbol{\beta}}) - \frac{A_i - \pi(\boldsymbol{X}_i, \hat{\gamma})}{1 - \pi(\boldsymbol{X}_i, \hat{\gamma})} \mu(0, \boldsymbol{X}_i, \hat{\boldsymbol{\beta}}). \tag{6}$$

Here, $\pi(\mathbf{X}, \hat{\gamma})$ denotes a model for the propensity score and $\mu(A, \mathbf{X}, \hat{\boldsymbol{\beta}})$ denotes a model for the outcome. This estimator satisfies a double robustness property, meaning that it remains consistent for $E\{Y^{g_T(\mathbf{X})}\}$ if either $\pi(\mathbf{X}, \hat{\gamma})$ or $\mu(A, \mathbf{X}, \hat{\boldsymbol{\beta}})$ is correctly specified. Note that in the case of an RCT with a balanced allocation design, $\pi(\mathbf{X}, \hat{\gamma}) = 1/\#(\text{treatment alternatives})$; alternatively, empirical proportions can be used to account for chance imbalance of treatment assignment conditional on the pretreatment characteristics [41]. In our simulation study we will use $\hat{C}_{AIPWE}(\mathbf{X})$ for estimating the contrast function (because of its double robustness property) and a parametric regression for the outcome model. Since we only consider RCT data in our study, the propensity model is always specified correctly; hence, because of double robustness, mis-specification of the outcome model $\mu(A, \mathbf{X}, \hat{\boldsymbol{\beta}})$ should not influence the performance of the method. Still, in our simulation study we want to investigate the possible effect of a mis-specification of the outcome model, firstly in order to set up a fair comparison of the methods under study, and secondly because double robustness is an asymptotic property only. Consequently, we will use two versions of the method of Zhang et al. [14] in our simulation study, namely one version with a correctly specified outcome model for estimating the contrasts and one with a mis-specified outcome model, where one of the relevant pretreatment characteristic has been dropped from the regression equation.

Once the contrasts are obtained, $g_T^{opt}(\mathbf{X})$ can be estimated by

$$\arg\max_{g_T \in G_T} [E\{g_T(\mathbf{X})\hat{C}(\mathbf{X})\}]. \tag{7}$$

Zhang et al. [14] were able to rewrite this as:

$$\arg\min_{g_T \in G_T} [E\{|\hat{C}(\mathbf{X})|[I\{\hat{C}(\mathbf{X}) > 0\} - g_T(\mathbf{X})]^2\}], \tag{8}$$

which makes it possible to recast the problem of estimating an optimal treatment regime as a weighted classification problem with $I\{\hat{C}(\mathbf{X}_i) > 0\}$ as the binary response, $\mathbf{X}_i$ as the predictor, $|\hat{C}(\mathbf{X}_i)|$ as the misclassification weight [42] and g_T as the classification rule looked for. Several existing classification techniques can be used to solve this classification problem, including Support Vector Machines, decision lists, and classification trees. Since we focus in the present paper on tree-based treatment regimes, we will use classification and regression trees (CART [43]) to estimate the treatment regime.

The recursive splitting procedure of CART stops if either the node becomes pure, that is, all cases in the node have identical values on the contrast or on all pretreatment characteristics, if the tree depth exceeds some predefined value, if the sample size in a node is less than a predefined minimum, if the split would result in a child node with a sample size less than a predefined child node size, or if for the best split of the node the improvement of the fit is smaller than some pre-specified minimum. The large initial tree will be pruned back to some optimal sub-tree using a form of cost-complexity pruning [39].

2.5 Qualitative interaction trees

The goal of QUINT is to identify subgroups that are involved in qualitative treatment-subgroup interactions, meaning that for some subgroups of patients, Treatment 0 is better than Treatment 1 whereas for other patients the reverse is true. This is done by partitioning the group of patients into two or three subgroups ($\wp_1$, $\wp_2$ and $\wp_3$), where each subgroup can comprise multiple leaves of a tree. For subgroup $\wp_1$ the outcome under Treatment 0 is higher than under Treatment 1, for subgroup $\wp_2$ the reverse is true, and for the optional subgroup $\wp_3$ the outcome is about equal under both treatments. QUINT tries to find the best partition of the sample of patients into these subgroups, in the sense that the qualitative treatment-subgroup interactions should have the largest possible practical significance. The latter means that the differences in outcome between Treatment 0 and Treatment 1 in the leaves assigned to $\wp_1$ and $\wp_2$ should be as large as possible, and that the numbers of patients assigned to $\wp_1$ and $\wp_2$ should be as large as possible. A treatment regime can be derived from the QUINT output based on the subgroups to which the leaves have been assigned, with for patients in leaves assigned to $\wp_1$ the optimal treatment being Treatment 0, and for patients in leaves assigned to $\wp_2$ the optimal treatment being Treatment 1. Obviously, for patients assigned to $\wp_3$ there is no clear optimal treatment alternative. However, to define an unambiguous treatment regime, we assign each leaf of $\wp_3$ to the treatment with the highest mean outcome in that leaf.

The partition criterion is a product of four components, namely the mean difference in outcome between Treatment 0 and Treatment 1 in the leaves assigned to $\wp_1$ and $\wp_2$ respectively, denoted by D_1 and D_2, and the number of patients assigned to $\wp_1$ and $\wp_2$, denoted by M_1 and M_2, respectively:

$$Q = w_1[\log(1 + D_1) + \log(1 + D_2)] + w_2[\log(M_1) + \log(M_2)], \tag{9}$$

where w_1 and w_2 are weights to put the difference in outcome and cardinality components on comparable measurement scales. After a check whether qualitative treatment-subgroup interactions are present, QUINT follows a stepwise binary splitting procedure which implies that in each step, a node, a pretreatment characteristic, a split of that characteristic, and an assignment of the terminal nodes or leaves of the current tree to $\wp_1$, $\wp_2$ (and possibly $\wp_3$) are chosen that maximize the partition criterion.

The algorithm stops splitting if a split does no longer imply a higher value of Q, if the number of leaves is greater than a predefined maximum, or if the sample size for one of the treatments in a leaf is lower than a predefined value. To avoid overfitting, the maximal tree is pruned back to some optimal sub-tree using a bootstrap-based procedure for bias correction [39, 44].

3 Method of the simulation study

In the simulation study, we will evaluate the performance and error rates of the four methods under study, and we will examine how these are influenced by certain characteristics of the data and the data-generating mechanism. We will first introduce the design of the simulation study, and then elaborate on the specification of the methods under comparison.

3.1 Design

For each dataset we created n observations $(Y_i, A_i, \mathbf{X}_i)$, $i = 1, ..., n$, where $\mathbf{X}_i = (X_{i1}, ..., X_{ij})^T$, $j = 1, ..., J$, was generated from a multivariate normal distribution with $\mu_{X_j} = 0$, $\sigma_{X_j} = 1$, and $\rho_{jj'} = 0.4$; A was generated from a Bernoulli distribution with $\theta = 0.50$; outcomes were generated as $Y_i = \mu(A_i, \mathbf{X}_i) + \epsilon_i$ with ϵ_i standard normal and with the link function being defined below.

The following factors were manipulated in the data generation process:

- Sample size, n with levels $n = 300$ and $n = 1000$
- The number of pretreatment characteristics J, with levels $J = 5$ and $J = 20$.
- The link function, with levels identity and exponential link, leading to the following conditional expectations $\mu(A, \mathbf{X})$:

$$\mu(A, \mathbf{X}) = 1.0 + 0.25X_1 + 0.25X_2 - 0.25X_5 - c[A - g^{opt}(\mathbf{X})]^2 \tag{10}$$

and

$$\mu(A, \mathbf{X}) = \exp\{1.0 + 0.25X_1 + 0.25X_2 - 0.25X_5 - c[A - g^{opt}(\mathbf{X})]^2\} \tag{11}$$

with g^{opt} denoting the true optimal treatment regime (defined below) and c being related to the cost implied by assigning the patient under study to the non-optimal treatment alternative.

- The effect size (Cohen's d [45]) of the difference in outcome between the optimal and non-optimal treatment alternative with levels medium ($d = 0.5$), and large ($d = 1$). These levels imply that under the identity link function, c is equal to 0.5 or 1, whereas under the exponential link function c is equal to 0.22 or 0.44.
- The true optimal treatment regime, g^{opt} with five levels, where Levels (1) and (2) refer to simple and complex tree-based scenarios, Levels (3) and (4) refer to simple and complex non-tree scenarios, and Level (5) to a quantitative scenario, in which the underlying regime is a trivial tree according to which the optimal alternative is Treatment 1 for every patient:

$$\begin{aligned}
&(1)\ g_T^{opt}(X) = I(X_1 > -0.545)I(X_2 < 0.545)\\
&(2)\ g_T^{opt}(X) = [I(X_1 < -0.545)I(X_2 > 0.545)] \vee [I(X_1 < -0.545)I(X_2 < 0.545)I(X_3 < 0.545)]\\
&(3)\ g^{opt}(X) = I(X_1 > X_2)\\
&(4)\ g^{opt}(X) = I(X_1 > (X_2^2 - 1)).
\end{aligned}$$

In the fifth scenario, the final term $-c[A - g^{opt}(\mathbf{X})]^2$ in eqs (10) and (11) was replaced by

$$(5)\ -0.25\{I(X_1 \le -0.43)[A - I(X_1 \le -0.43)]^2\} - c\{I(X_1 > -0.43)[A - I(X_1 > -0.43)]^2\},$$

with 0.25 now relating to the cost implied by assigning the patient under study to the non-optimal treatment alternative when $X_1 \leq -0.43$ and c to the cost implied when assigning the patient under study to the non-optimal treatment alternative when $X_1 > -0.43$.

All design factors were fully crossed, yielding 2 (sample size) × 2 (number of pretreatment characteristics) × 2 (link function) × 2 (effect size) × 5 (optimal treatment regime) = 80 combinations, with 500 Monte Carlo replications for each combination, yielding 40,000 data sets in total.

3.2 Specification of the methods

In this section, we will explain how we specified the options and tuning parameters for the four methods under study, in such a way that the comparison of these methods will be as fair as possible.

3.2.1 Model-based recursive partitioning

The R-package `party` [46] was used to estimate tree-based treatment regimes with MOB, with $E(Y|X, A) = \beta_0 + \beta_1 A$ as model[2], with the pretreatment characteristics $X_1, ..., X_J$ as possible split variables, and with default settings for all tuning parameters, except that the minimum number of persons in a node to split was set at 40 (`minsplit = 40`).

3.2.2 Interaction trees

IT was applied using R-code provided by the author. For λ, a penalty parameter used in the pruning procedure, we used ln(n); the minimum number of persons in a node to split was set at 40, the minimum number of persons in a terminal node at 20, and the maximum depth of the tree at six. We used two thirds of the data as learning sample, and one third as test sample.

3.2.3 Optimal treatment regime identification by Zhang et al.

We used self-built R-code for the contrast calculation. For this calculation we used as outcome model on the one hand $\mu(A, \boldsymbol{X}, \hat{\boldsymbol{\beta}}) = (\beta_0 + \sum_{j=1}^{J} \beta_j X_j) + A(\beta_0^* + \sum_{j=1}^{J} \beta_{J+j} X_j)$[3]. On the other hand, we also used as outcome model the same regression model from which one relevant predictor (X_1) was dropped from both the main effect and the interaction term: $\mu(A, \boldsymbol{X}, \hat{\boldsymbol{\beta}}) = (\beta_0 + \sum_{j=2}^{J} \beta_j X_j) + A(\beta_0^* + \sum_{j=2}^{J} \beta_{J+j} X_j)$. With regard to the propensity model, instead of fixing it to 0.5, we used empirical proportions to account for chance imbalance of treatment assignment. We estimated the contrasts using $\hat{C}_{AIPWE}(\mathbf{X}_i, A_i, Y_i, \hat{\boldsymbol{\beta}})$ and defined binary responses $\hat{Z}_i = I(\hat{C}_{AIPWE}(\mathbf{X}_i, A_i, Y_i; \hat{\boldsymbol{\beta}}) > 0)$, and weights $\hat{W}_i = |\hat{C}_{AIPWE}(\mathbf{X}_i, A_i, Y_i; \hat{\boldsymbol{\beta}})|$. The new dataset $\{\hat{Z}_i, \mathbf{X}_i, \hat{W}_i\}$ was analyzed using `rpart` [47] with default settings for all tuning parameters, except that we used $\hat{W}$ as the individual misclassification costs (`weights=` $\hat{W}$), the minimum number of persons in a node to split was set at 40 (`minsplit = 40`), the minimum number of persons in a terminal node at 20 (`minbucket =`

2 Note that for some conditions this model is mis-specified with regard to the link function as in the data generation process for half of the conditions an exponential link function was used (see eq. (11)) whereas the outcome model always included an identity link.

3 The main effects and interactions with treatment of all available (rather than all relevant) predictors were included in the outcome model. Also note that this outcome model is mis-specified with regard to the interaction terms, as in the data generating process these were based on rectangular regions (see Section 10.3). On top of that, as was the case for MOB, the outcome model was mis-specified in terms of the link function for conditions in which an exponential link function was used in the data generating process.

20), and the maximum depth of the tree at six (maxdepth = 6). For the complexity parameter used in the pruning procedure (cp), we used the complexity parameter associated with the smallest cross-validated error [48].

3.2.4 Qualitative interaction trees

To estimate a tree-based treatment regime using QUINT, we used the R-package quint [49] with default settings for all tuning parameters, except that we set the minimum number of persons with respectively $A = 0$ and $A = 1$ in a leaf at 10 (a1=a2=10).

4 Results

The goal of the simulation study was to compare the performance of the four methods in an absolute and a relative sense. A treatment regime was estimated for each dataset using each of the four methods; subsequently, we performed a repeated measures analysis of variance (ANOVA) for each outcome measure with method as a within-subjects variable and the manipulated factors as between-subjects variables. We will report all effects (including interactions) with an effect size higher than some cut-off value. We will use η^2 as effect size, where $\eta^2 = \frac{SS_{effect}}{SS_{total}}$. Here, SS_{effect} is the sum of squares of the effect and SS_{total} are the total sum of squares. We calculated SS_{total} separately for the between- and the within-subject effects. This implies that the η^2-values of all between-subject effects and the interactions between these effects sum up to 1; similarly, the η^2-values for the within-subject effect and the interactions between the within- and between-subject effects sum up to 1. As a cut-off, we used 0.10, implying that an effect was reported if it explained more than 10 % of the corresponding total sum of squares.

4.1 Expected outcome of the regime in the population

The first evaluation criterion pertained to the expected outcome if all members of the population under study would be assigned to the treatment alternative implied by the regime under study, $E(Y^{\hat{g}_T^{opt}})$. To approximately calculate $E(Y^{\hat{g}_T^{opt}})$, we used a Monte Carlo averaging procedure proposed by Zhang et al. [14]: We simulated a super-sample, that is, a set of $N = 10^6$ patients on the basis of the true data generating model, subjected this to the estimated treatment regime, and calculated the average of the resulting outcomes under this regime. Yet, unfortunately, the measure $E(Y^{\hat{g}_T^{opt}})$ suffers from the problem that its values are not comparable for data simulated in different conditions, for two reasons. First, the maximally possible value of $E(Y^{\hat{g}_T^{opt}})$ is $E(Y^{g^{opt}})$, which differs substantially between conditions. Second, in each condition one may consider as possible $\hat{g}_T^{opt}$ a trivial regime according to which all patients should be assigned to the marginally best treatment alternative ($g = a^{opt}$). The performance of such a trivial regime ($E(Y^{g=a^{opt}})$) differs substantially between conditions, which implies that the difficulty of finding a regime $\hat{g}_T^{opt}$ with a better performance ($E(Y^{\hat{g}_T^{opt}})$) than the trivial regime differs between conditions as well. We accounted for these issues by introducing a measure of Normalized Performance Gain (NPG):

$$\frac{E(Y^{\hat{g}_T^{opt}}) - E(Y^{g=a^{opt}})}{E(Y^{g^{opt}}) - E(Y^{g=a^{opt}})}, \tag{12}$$

with $E(Y^{\hat{g}_T^{opt}}) = \frac{1}{N}\sum_{i=1}^{N}[y_i^1 \cdot \hat{g}_T^{opt}(x_i) + y_i^0 \cdot (1 - \hat{g}_T^{opt}(x_i))]$, $E(Y^{g^{opt}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i^1 \cdot g^{opt}(x_i) + y_i^0 \cdot [1 - g^{opt}(x_i)])$, and $E(Y^{g=a^{opt}}) = \frac{1}{N}\sum_{i=1}^{N}(y_i^1 \cdot a^{opt} + y_i^0 \cdot [1 - a^{opt}])$. The NPG (with a maximum value of 1 in every condition) indicates the fraction of the theoretically possible performance gain that is achieved by assigning all patients in the population to a treatment alternative based on a regime $\hat{g}_T^{opt}$ as compared to the trivial regime of assigning

Table 1: Effects with $\eta^2 > 0.10$ resulting from the ANOVA on the normalized performance gain (NPG).

Effect	η^2
Sample size	0.17
Effect size	0.25
Scenario	0.26
Method	0.30

all patients to the marginally best treatment regime $g = a^{opt}$. A limitation of this measure is that it can only be used if the denominator in eq. (12) is different from 0, that is, if the true optimal regime g^{opt} is non-trivial. As such, the NPG cannot be used for the fifth scenario with a quantitative treatment-subgroup interaction. However, for this scenario the Type I error rate is the primary outcome of interest. For all other scenarios, the NPG was analyzed as a first outcome measure.

The grand mean of the NPG was 0.542. The ANOVA results of this outcome measure revealed that it was mainly influenced by *sample size*, *effect size*, *scenario*, and *method* (see Table 1). The main effects of sample size and effect size read that, as expected, the NPG increases for larger sample sizes, and for larger effect sizes. The main effect of scenario reads that, in line with our expectations, the NPG was better under tree-based scenarios than under non tree-based scenarios and better under simple as compared to complex scenarios. Regarding the main effect of method: The approach of Zhang et al. [14], even with one omitted predictor, had the highest mean NPG, followed by MOB; the mean NPGs of QUINT and IT were lower. These marginal means are shown in Table 5.

4.2 Proportion of patients assigned to the truly best treatment alternative

We calculated for every data set the proportion of patients who were assigned by the estimated treatment regime, $\hat{g}_T^{opt}$, to the truly best treatment alternative according to the true optimal treatment regime underlying the data, g^{opt}. The grand mean of the proportion of patients assigned to the truly best treatment alternative was 0.856. An ANOVA revealed that this proportion was mainly influenced by *effect size*, *scenario*, *method*, and a *method*scenario* interaction (see Table 2). For effect size, we observed the same trend as for the NPG, namely that the proportion of correctly assigned patients increases for larger effect sizes. Regarding the main effect of scenario the proportion of correctly assigned patients was higher in tree-based than non tree-based scenarios, in line with our expectations; yet, somewhat unexpectedly, the proportion of correctly assigned patients was also slightly lower in the simple than in the complex scenarios. We will go into this in more depth in the Discussion Section. Regarding the effect of method, the proportion of correctly assigned patients was highest for the method of Zhang et al. [14] without and with one omitted predictor, followed by MOB, IT, and QUINT. These marginal mean proportions are shown in Table 5. Finally, the interaction between method and scenario reads that IT and QUINT have a pattern that differs from that of the other methods, with more

Table 2: Effects with $\eta^2 > 0.10$ resulting from the ANOVA on the proportion of patients classified to the correct treatment alternative.

Effect	η^2
Effect size	0.12
Scenario	0.56
Method	0.26
Method*scenario	0.15

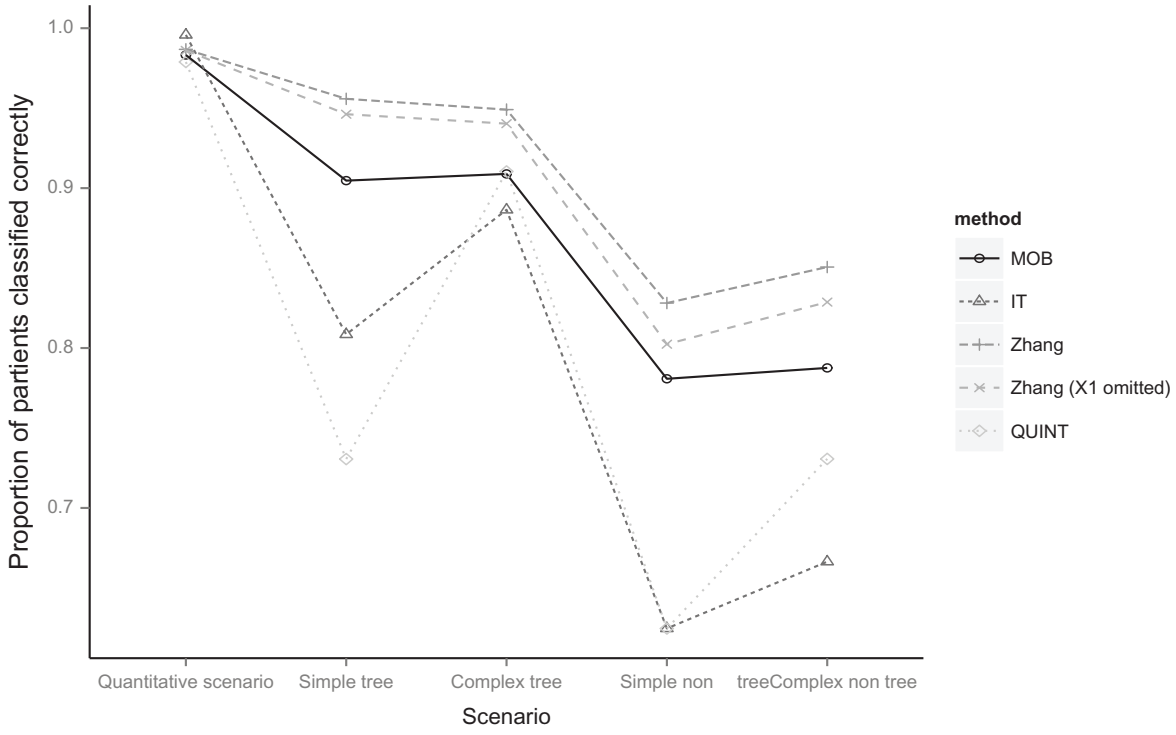


Figure 2: *Method*scenario* interaction for the proportion of patients assigned to the truly best treatment by the estimated treatment regimes (where Zhang (X1 omitted) refers to the method of Zhang et al. [14] where X_1 was omitted from the regression equation).

pronounced differences in the proportion of correctly assigned patients between the tree-based and non tree-based scenarios, and between the simple and complex scenarios. This interaction is depicted in Figure 2.

4.3 Type I error rate

The quantitative interaction scenario was used to evaluate the Type I error rate. A Type I error is committed when the null hypothesis that everyone should be assigned to the same treatment alternative is wrongly "rejected", meaning that a non-trivial tree in which the leaves are assigned to different treatments, is returned. For each cell of the design under the fifth scenario, which involves a quantitative treatment-subgroup interaction only, we calculated for each method the proportion of data sets for which a non-trivial tree with the leaves assigned to different treatment alternatives, was returned. The resulting proportions were subjected to an ANOVA, with the highest order interaction being used as error term.

Table 3: Effects with $\eta^2 > .10$ resulting from the ANOVA on the Type I error rates.

Effect	η^2
Link function	0.93
Method	0.38
Method*number of pretreatment characteristics	0.13
Method*link function	0.13
Method*effect size	0.18

The grand mean of the Type I error rate was 0.084. The ANOVA revealed that this error rate was primarily influenced by *link function* and *method* main effects, and by *method*link function* and *method*effect size* interactions (see Table 3). Regarding the main effect of link function: The Type I error probability was lower for an exponential than for an identity link function; yet, from the interaction between method and link function we can conclude that for IT, unlike for the other methods, the Type I error rate is not influenced by the link function. The main effect of method implies that for IT the probability of making a Type I error was smallest, followed by MOB, the method of Zhang et al. [14] without and with an omitted predictor, and QUINT. These marginal mean error rates are shown in Table 5. Note that for MOB and IT, only non-trivial trees in which the leaves are assigned to different treatment alternatives are being counted as a Type I error. Regarding the interaction between method and the number of pretreatment characteristics: Only for QUINT there was an effect of the number of pretreatment characteristics on the Type I error probability, with a lower Type I error probability in case of five as compared to 20 pretreatment characteristics. Similarly, the interaction between method and effect size reads that only for QUINT a higher effect size implied a lower Type I error probability. Obviously, these interactions are mainly based on deviating patterns of two backbench methods (similar to the interaction illustrated before in Figure 2); as these are not of primary interest to answering our research questions, we omit figures to graphically represent them.

4.4 Type II error rate

The Type II error, defined as the probability that the null hypothesis (that everyone should receive the same treatment alternative) is incorrectly 'retained', meaning that a trivial tree or a non-trivial tree in which all leaves are assigned to the same treatment alternative, will be estimated using the simple and complex tree-based and non tree-based scenarios. For every cell of the design under these scenarios, we calculated the proportion of data sets for which a trivial tree or a non-trivial tree with all leaves assigned to the same treatment alternative, was returned. Again, the highest order interaction was used as error term in the ANOVA.

The grand mean of the Type II error rate was 0.185. An ANOVA revealed that the Type II error rate was primarily influenced by *scenario*, *sample size*, *effect size*, and *method* main effects, as well as by *method*sample size*, and *method*scenario* interactions (see Table 4). Regarding the effect of scenario, the Type II error rate was higher under the simple than under the complex scenarios, and higher under the non tree-based than under the tree-based scenarios; however, the interaction between method and scenario indicated that only for QUINT the difference between simple and complex scenarios was large, and especially for IT the difference between tree-based and non tree-based scenarios was large, whereas for the other methods the Type II error remained about equal over the different scenarios. The main effects of sample size and effect size were in the expected direction: The Type II error rate was lower for larger sample sizes and for larger effect sizes. The main effect of method lined up with that of the first two performance measures: The method of Zhang et al. [14] without one omitted predictor had the lowest Type II error probability, followed by the same method with omitted predictor and MOB; QUINT and IT had a higher probability of making

Table 4: Effects with $\eta^2 > .10$ resulting from the ANOVA on the Type II error rates.

Effect	η^2
Scenario	0.16
Sample size	0.16
Effect size	0.49
Method	0.55
Method*sample size	0.13
Method*scenario	0.14

Table 5: Marginal means of the four different outcome variables for the values of the independent variables with a main effect with $\eta^2 > 0.10$ (in case of a main effect with $\eta^2 < 0.10$, the corresponding cells were left blanc).

Effect	NPG	Proportion	Type I error	Type II error
Sample size				
300	0.45			0.33
1000	0.65			0.13
Effect size				
0.5	0.41	0.81		0.28
1.0	0.68	0.90		0.09
Link function				
Identity			0.13	
Exponential			0.04	
Scenario				
Quantitative scenario	NA	0.99	0.08	NA
Simple tree	0.70	0.87	NA	0.18
Complex tree	0.65	0.92	NA	0.10
Simple non-tree	0.43	0.73	NA	0.25
Complex non-tree	0.39	0.77	NA	0.21
Method				
MOB	0.58	0.87	0.09	0.06
IT	0.37	0.79	0.01	0.45
Zhang	0.71	0.91	0.10	0.03
Zhang (X_1 omitted)	0.67	0.90	0.10	0.04
QUINT	0.48	0.80	0.12	0.38

Type II errors. These marginal mean Type II error rates are shown in Table 5. Regarding the interaction effect between method and sample size: The pattern of IT is different from that of the other methods, with a more pronounced difference in Type II error rate between the two sample sizes. The pattern of QUINT is opposite to that of the other methods, with a larger Type II error rate for the larger sample size.

5 Illustrative application

We estimated a tree-based treatment regime with each of the four methods under study using a Clinical Trials Network data set on the evaluation of integrating motivational interviewing techniques into the initial contact and evaluation sessions of behavior therapies [50]. This RCT data set consists of 423 individuals seeking treatment for a substance use problem. Participants were randomly assigned to standard intervention (n = 214) or standard intervention with integrated motivational interviewing (MI) techniques (n = 209). The dataset includes 18 pretreatment characteristics, among which demographical characteristics (e.g., age, gender, ethnicity) and aspects of substance use (e.g., composite scores of the Addiction Severity Index (ASI), see McLellan et al. [51]). The outcome measure on which we will focus here, is the total number of days without substance use during the 28 days after randomization. The original analyses of these data revealed no significant effect of treatment on the number of days without substance use for the sample as a whole. However, in the discussion, the authors hint at possible treatment effect heterogeneity and state that it is important to understand the types of individuals for whom MI is effective and what mediators and moderators might impact the process.

We excluded all observations with missing values on the outcome of interest (n= 72). We used the same settings, tuning parameters and code as in the simulation study, except that for IT we took the whole data set as learning sample and used a bootstrap procedure for pruning, as we learned from the simulation study that a large sample size is of utmost importance for the performance of IT. Also, for QUINT we used the difference in means criterion instead of the effect size, since the authors advise to do so when the outcome is

measured on a scale with values that have a clear pragmatic meaning [13]; besides we used 0.20 as minimum absolute value of the standardized mean difference in treatment outcome in the two leaves for a tree to be grown (`dmin`), since we learned from the simulation study that the risk of a Type II error is very high using `dmin=0.30`.[4]

The tree-based treatment regimes estimated with the methods under study, are shown in Figure 3. All methods return a non-trivial tree with the leaves being assigned to different treatments. For three of the methods, employment is an important splitting variable with 0.64 as cut-off, where persons with an employment score > 0.64 benefit from standard treatment. For QUINT and IT more than one split is based on employment, hinting at a possible curvilinear effect of this variable. Finally, both QUINT and IT split based on years of education as well, with a cut-off value of about 12 years. Here, persons with more than 12 years of education and a low employment score should receive standard treatment, whereas persons with less than 12 years of education and a low employment score should receive MI. MOB uses the number of days that a patient has used drugs during 30 days prior to treatment assignment as split variable. Although this split was mainly caused by a main effect of this variable (with persons who have used drugs during more than 19 days out of 30 days prior to treatment assignment having a lower intercept), the treatment effect, although small, had a different sign in both leaves as well, implying a different optimal treatment alternative in each leaf.

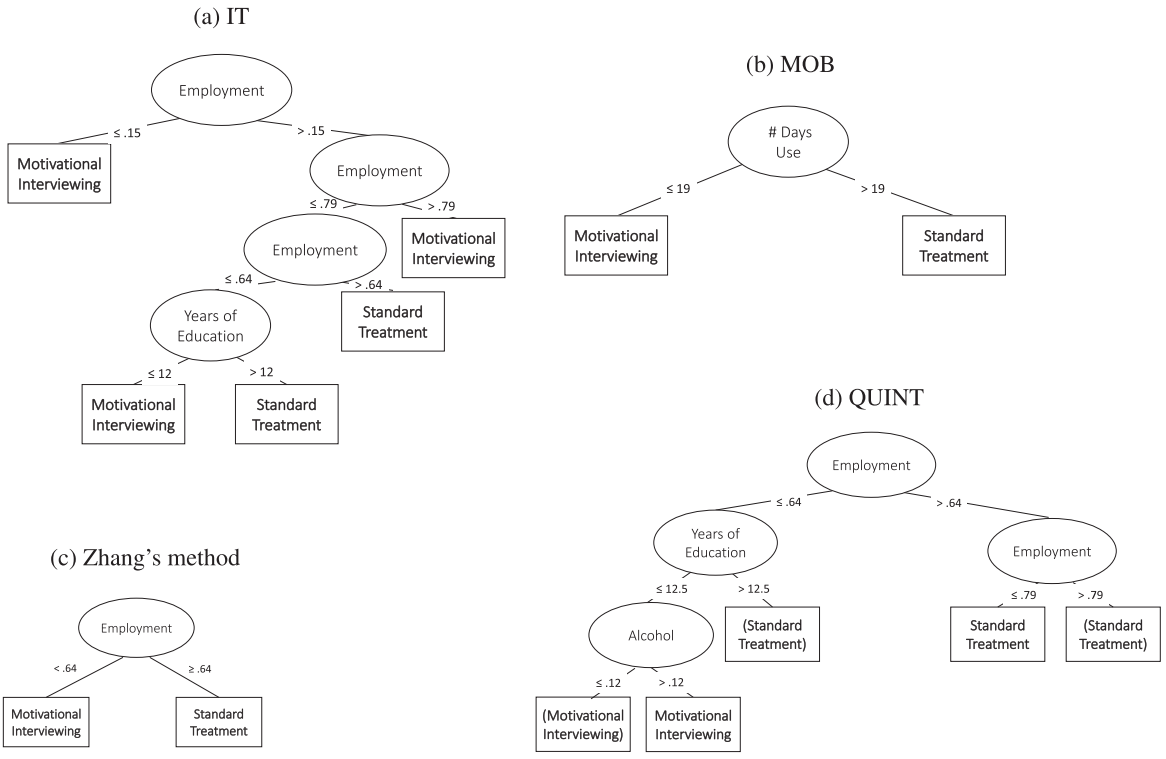


Figure 3: Tree-based treatment regimes estimated from the data set of Caroll et al. (2006) [50] on the evaluation of integrating motivational interviewing techniques into the initial contact and evaluation sessions of behavior therapies. Each leaf is associated with an optimal treatment. In the QUINT solution the leaves with "no difference" were assigned to the treatment alternative (displayed between parentheses) with the highest mean outcome in that leaf. A higher score on the ASI variables (e.g., employment, alcohol, psychological) indicates more problems in that area. # Days use indicates the number of days the patient had used drugs during 30 days prior to treatment assignment.

4 As the outcome of interest is the number of days without substance use, a model for count data would be appropriate here. We analyzed the data for the methods that can deal with such data (viz., MOB and Zhang's method) again using a Poisson regression model. The resulting regimes remained the same.

One could base the choice between the four treatment regimes in Figure 3 on an estimate of the expected outcome of the regimes in the population ($\hat{E}[\hat{g}_T^{opt}]$), using a naive estimator (where the mean outcome under the optimal treatment alternative in a leaf is multiplied by the number of persons in that leaf) or using the (doubly robust) AIPW estimator [40, 9]. However, for standard usage of these estimators, the same data are used for estimating a regime and evaluating it, leading to positively biased results, with this bias most probably being higher for larger trees. Indeed, when evaluating the tree-based treatment regimes using the AIPWE, the largest tree had the highest estimated expected outcome (IT: 26.50), followed by the second largest tree (QUINT: 25.78), and finally the smaller trees (Zhang: 25.49 and MOB: 25.14). Similar results were obtained using the naive estimator.[5]

6 Discussion

We evaluated the absolute and relative performance of four methods for estimating tree-based treatment regimes (viz., Interaction Trees, Model-based Recursive Partitioning, Qualitative Interaction Trees and two implementations of the method of Zhang et al. [14]) in terms of the expected outcome in the population if all patients would receive treatment according to the estimated treatment regime and in terms of the proportion of patients assigned to the truly best treatment alternative. Besides, we compared the methods in terms of their probability of making inferential errors (viz., Type I and Type II errors). We will now discuss the major findings of our simulation study.

To start with, as a background issue, we were interested in how certain data characteristics would influence the performance of the whole of the methods under study. We found that under larger sample sizes and effect sizes, the methods achieve higher expected outcomes and higher proportions of patients assigned to the truly best treatment alternative, as well as lower Type II error rates. Also, as expected, we found that the methods perform better under tree-based than under non tree-based scenarios in terms of the expected outcome and the proportion of patients assigned to the truly best treatment alternative. Furthermore, the expected outcome was higher in simple than complex scenarios. Somewhat unexpectedly, the latter did not hold for the proportion of correctly assigned patients. This can be explained by the higher Type II error probabilities in the simple scenarios, which can be attributed to the greedy nature of tree algorithms (meaning that a split will only be made if there is an immediate advantage of it). For example, in the context of estimating optimal tree-based treatment regimes, this means that a split is only being made if in one of the daughter nodes the expected potential outcome of Treatment 1 clearly exceeds that of Treatment 0, whereas in the other daughter node the reverse holds true. This implies a qualitative treatment-split interaction. For the first split in the complex scenario, this qualitative treatment-split interaction was more pronounced than for the first split in the simple scenario. On top of this, in the simple scenarios there were more patients for whom the optimal treatment alternative was not the marginally best treatment alternative, implying that in the case of a trivial tree (or a non-trivial tree with all leaves assigned to the same treatment alternative) being returned, more patients were assigned to the incorrect treatment alternative. Type I error rates were higher under an identity than under an exponential link function. The latter can be explained by the fact that the absolute difference between the two treatments in the leaves is higher under the exponential link function, making it easier to recognize whether a qualitative treatment-subgroup interaction is present or not. Finally, quite unexpectedly, only for QUINT, the Type I error rate increased when the number of pretreatment characteristics increased. The fact that MOB is resistant to an increase of the number of pretreatment characteristics may be explained at least in part by the internal stopping criterion of MOB, which is based on a significance threshold that takes into account a multiple comparison correction. For IT, the lack of increase of the Type

5 Optionally, on top of the optimal treatment alternative, one could be interested in the size of the relative treatment effect for the induced subgroups of patients, for example, based on the fitted model or based on a post-hoc estimation procedure (e.g., making use of the AIPWE). For reasons of parsimony we did not include this in the present application.

I error risk may be explained by a floor effect for the Type I error risk, probably due to the use of a strong penalty in the pruning procedure.

The primary question we wanted to answer in this paper was of course which method performs best (under which circumstances). We learned that the method of Zhang et al. [14], even with one omitted predictor, performs best across almost all conditions of our simulation study, in terms of the expected outcome, the proportion of patients assigned to the truly best treatment alternative, and the Type II error rate. In terms of the same outcome measures, the performance of MOB was usually somewhat lower than that of the method of Zhang et al. [14], whereas the performance of QUINT and IT was considerably lower in most conditions of the simulation study. However, in terms of the Type I error rate, the picture was quite different from that of the other outcome measures: Here, the performance of IT was best (with a very low Type I error rate), that of MOB was somewhat in between, whereas QUINT and the method of Zhang et al. [14] had the highest Type I error rates. Besides, although we found some interactions between method and data characteristics, the positions of the best performing methods with respect to the different outcome measures were in general not influenced by those data characteristics.

Two remarks with regard to these findings, seem to be in place. First, we implemented MOB as suggested by Zeileis et al. [52] and Seibold et al. [36], using a simple regression model ($E(Y|X, A) = \beta_0 + \beta_1 Treatment$) and an instability check of the parameters (β_0, β_1) in each node. One might wonder what would happen if one would expand this regression model by adding main effects of the predictors (resulting into $\beta_0 + \beta_1 Treatment + \sum \beta_{j+1} X_j$), and performing an instability check on the full set of β-parameters. We repeated some cells of the simulation design for MOB in this way, which led to changeable results, with a considerable deterioration of NPG in some conditions, a moderate improvement in other conditions, and the NPG staying the same in still other conditions (see Supplementary Materials, Table 1). Additional research is needed to further characterize the effect of implementing MOB in this way. The same holds true for the recently proposed implementation of the MOB approach by Seibold et al. [53]. They suggest to include the main effects of the predictors in the model, while keeping their effects constant over the subgroups and therefore examining the instability of (β_0, β_1) only. Using MOB in this way could lead to better results when there are prognostic variables that influence the outcome in a linear way (as was the case in our simulation study), because the regression model will account for these main effects of the predictors and MOB will no longer try to approximate them by step-functions.

Second, it is not surprising that the method of Zhang et al. [14] performs best with regard to the NPG, since it was the only method in our comparison that maximizes an estimate of the expected outcome of a regime. To get an idea of the performance of the method of Zhang et al. [14] relative to a method that also optimizes such an estimate, we compared its performance to the performance of the method of Zhang et al. [21]. This method also estimates optimal treatment regimes based on rectangular regions, yet it represents the treatment regime by a decision list instead of a decision tree; we will further refer to this method as the decision list method. We applied the decision list method with default settings to part of the data sets generated in the original simulation study. In general, in terms of the NPG, the performance of the two methods was similar. To be sure, the decision list method slightly outperformed the method of Zhang et al. [14] in conditions with a small sample size or effect size, whereas in conditions with a large sample size and effect size, the reverse held true. With regard to Type II errors, the performance of the decision list method turned out to be perfect, whereas for the method of Zhang et al. [14], non-negligible Type II error rates were found in conditions with a small sample size and effect size. Yet, the price to be paid for the perfect Type II error performance of the decision list method is substantial, as the Type I error risk of this method turned out to be very high. Often, it was more than twice as high as the Type I error risk of the method of Zhang et al. [14]. Detailed results on this are reported in Tables 2 and 3 of the Supplementary Materials.

Summarizing, the most important result we found is that the method of Zhang et al. [14] is superior in terms of the expected outcome, the proportion of patients assigned to the truly best treatment alternative and the Type II error rate as compared to the other tree-based methods (in terms of the same outcome measures, the performance of the decision list method was slightly better than that of the method of Zhang et al. [14]); yet, the method of Zhang et al. [14] performs relatively bad in terms of the Type I error rate (and the

performance of the decision list method was even worse in terms of this outcome measure). One may obviously wonder whether these two seemingly contradictory observations can be reconciled. In fact, committing a Type I error does not necessarily imply a considerably lower expected outcome or proportion of correctly assigned patients. Indeed, when looking at the output of the method of Zhang et al. [14] in more detail, when a Type I error is committed, usually a large tree is grown, from which only a small group of patients is assigned to the incorrect treatment alternative; if, on top of that, for this group of patients the difference in outcome between both treatment alternatives is small, the influence on the expected outcome will be negligible. On the other hand, making a Type II error may lower the expected outcome and proportion of correctly assigned patients considerably, since this implies that a fairly large group of patients is assigned to the incorrect treatment alternative. Note that all this could further imply that the design of our simulation study, with many scenarios involving a qualitative treatment-subgroup interaction (with a risk of committing a Type II error), and only one scenario involving a quantitative treatment-subgroup interaction (with a risk of committing a Type I error) might have given the liberal method of Zhang an advantage over more conservative methods.

The answer to the question which method to use, might at first sight seem straightforward as the method of Zhang et al. [14] (or the decision list method if one is interested in regimes represented by decision lists) performs best on three out of four outcome measures. Indeed, especially if one is primarily interested in estimating a treatment regime that maximizes the mean outcome in the population, this method looks like the best choice. Yet, one should perhaps also not overlook the high Type I error rate of the method of Zhang et al. [14] (and the decision list method). Such a high Type I error rate implies a high risk of identifying non-trivial treatment regimes that cannot be replicated in follow-up research. Problems of non-replicability recently have been denounced in the area of treatment-subgroup interactions, which immediately links up with the area of non-trivial treatment regimes (see e.g., Rothwell [28], Assmann et al. [29], Brookes et al. [32], Wang et al. [30], Feinstein [33]). One would probably care more about problems of non-replicability, if, beyond a search for actuarial treatment assignment rules and beyond the aim of maximizing the mean outcome of a treatment regime in the population, one would also like to get some insight into the underlying decision structure of the regime, and especially into the why of this structure (e.g., in terms of the underlying mechanisms of recovery processes and of effective treatment components). At this point, it seems fair to conclude that the method of choice should depend on one's objectives. Whereas the regimes estimated by the method of Zhang et al. [14] (and the decision list method) might result in the highest expected outcome in the population, if one cares about avoiding non-replicable results or about drawing correct conclusions regarding recovery mechanisms or effective treatment components, at least supplemental analyses with the fairly liberal method of Zhang et al. [14] (or the decision list method) by analyses with a more conservative method might be preferable.

Funding: The research reported in this paper was supported in part by the Research Fund of KU Leuven (GOA/15/003), and by the Interuniversity Attraction Poles programme financed by the Belgian government (IAP/P7/06). The computational resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation – Flanders (FWO) and the Flemish Government – department EWI. The information reported in the illustrative application results from secondary analyses of data from clinical trials conducted by the 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. NIDA databases and information are available at http://datashare.nida.nih.gov.

References

1. Dehejia RH. Program evaluation as a decision problem. J Econometrics 2005;125:141–73.
2. Heckman JJ. Micro data, heterogeneity, and the evaluation of public policy: nobel lecture. J Political Econ 2001;109:673–748.
3. Hamburg MA, Collins FS. The path to personalized medicine. New Eng J Med 2010;363:301–4.

4. Lipkovich I, Dmitrienko A, D'Agostino Sr RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. Stat Med 2017;36:136–96.
5. Doove L, Dusseldorp E, Van Deun K, Van Mechelen I. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment–subgroup interactions. Adv Data Anal Classification 2014;8:403–25.
6. Murphy SA. Optimal dynamic treatment regimes. J R Stat Soc Ser B (Stat Methodol) 2003;65:331–55.
7. Zhang B, Tsiatis AA, Laber EB, Davidian M. Robust estimation of optimal dynamic treatment regimes for sequential treatment decisions. Biometrika 2013;100:581–694.
8. Moodie EE, Richardson TS, Stephens DA. Demystifying optimal dynamic treatment regimes. Biometrics 2007;63:447–55.
9. Zhang B, Tsiatis AA, Laber EB, Davidian M. A robust method for estimating optimal treatment regimes. Biometrics 2012;68:1010–8.
10. Qian M, Murphy SA. Performance guarantees for individualized treatment rules. Ann Stat 2011;39:1180.
11. Brinkley J, Tsiatis A, Anstrom KJ. A generalized estimator of the attributable benefit of an optimal treatment regime. Biometrics 2010;66:512–22.
12. Xu Y, Yu M, Zhao Y, Li Q, Wang S, Shao J. Regularized outcome weighted subgroup identification for differential treatment effects. Biometrics 2015;71:645–53.
13. Dusseldorp E, Van Mechelen I. Qualitative interaction trees: A tool to identify qualitative treatment–subgroup interactions. Stat Med 2014;33:219–37.
14. Zhang B, Tsiatis AA, Davidian M, Zhang M, Laber E. Estimating optimal treatment regimes from a classification perspective. Stat 2012a;1:103–14.
15. Laber E, Zhao Y. Tree-based methods for individualized treatment regimes. Biometrika 2015;102:501–14.
16. Su X, Tsai CL, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. J Mach Learn Res 2009;10:141–58.
17. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. J Comput Graph Stat 2008;17:492–514.
18. Foster J, Taylor J, Ruberg S. Subgroup identification from randomized clinical trial data. Stat Med 2011;30:2867–80.
19. Loh W, He X, Man M. A regression tree approach to identifying subgroups with differential treatment effects. Stat Med 2015;34:1818–33.
20. Su X, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. Int J Biostat 2008;4:Article 2. DOI: 10.2202/1557-4679.1071.
21. Zhang Y, Laber EB, Tsiatis A, Davidian M. Using decision lists to construct interpretable and parsimonious treatment regimes. Biometrics 2015;71:895–904.
22. Foster J, Taylor J, Kaciroti N, Nan B. Simple subgroup approximations to optimal treatment regimes from randomized clinical trial data. Biostatistics 2014;16:368–82.
23. Fu H, Zhou J, Faries DE. Estimating optimal treatment regimes via subgroup identification in randomized control trials and observational studies. Stat Med 2016;35:3285–302.
24. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search – a recursive partitioning method for establishing response to treatment in patient subpopulations. Stat Med 2011;30:2601–21.
25. Berger JO, Wang X, Shen L. A Bayesian approach to subgroup identification. J Biopharm Stat 2014;24:110–29.
26. Zhao Y, Zeng D, Rush AJ, Kosorok MR. Estimating individualized treatment rules using outcome weighted learning. J Am Stat Assoc 2012;107:1106–18.
27. Imai K, Ratkovic M, Estimating treatment effect heterogeneity in randomized program evaluation. Ann Appl Stat 2013;7: 443–470.
28. Rothwell PM. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. Lancet 2005;365:176–86.
29. Assmann SF, Pocock SJ, Enos LE, Kasten LE. Subgroup analysis and other (mis) uses of baseline data in clinical trials. Lancet 2000;355:1064–9.
30. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine – reporting of subgroup analyses in clinical trials. New Engl J Med 2007;357:2189–94.
31. Shaffer JP. Probability of directional errors with disordinal (qualitative) interaction. Psychometrika 1991;56:29–38.
32. Brookes ST, Whitley E, Peters TJ, Mulheran PA, Egger M, Davey Smith G. Subgroup analyses in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. Health Technol Assess 2001;5:1–56.
33. Feinstein AR. The problem of cogent subgroups: A clinicostatistical tragedy. J Clin Epidemiol 1998;51:297–9.
34. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Edu Psychol 1974;66:688–701.
35. Morgan SL, Winship C. Counterfactuals and causal inference. Cambridge: Cambridge University Press, 2014.
36. Seibold H, Zeileis A, Hothorn T. Model-based recursive partitioning for subgroup analyses. Int J Biostat 2016;12:45–63.
37. Zeileis A, Hornik K. Generalized m-fluctuation tests for parameter instability. Stat Neerlandica 2007;61:488–508.
38. Byar DP. Assessing apparent treatment – covariate interactions in randomized clinical trials. Stat Med 1985;4:255–63.
39. Leblanc M, Crowley J. Survival trees by goodness of split. J Am Stat Assoc 1993;88:457–67.
40. Bang H, Robins JM. Doubly robust estimation in missing data and causal inference models. Biometrics 2005;61:962–73.

41. Williamson EJ, Forbes A, White IR. Variance reduction in randomised trials by inverse probability weighting using the propensity score. Stat Med 2014;33:721–37.
42. Hastie TJ, Tibshirani RJ, Friedman JH. The elements of statistical learning: Data mining, inference, and prediction. New York: Springer, 2009.
43. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees. Belmont: Wadsworth, 1984.
44. Efron B. Estimating the error rate of a prediction rule: Improvement on cross-validation. J Am Stat Assoc 1983;78:316–31.
45. Cohen J. Statistical power analysis for the behavioral sciences. New York: Academic press, 2013.
46. Hothorn T, Hornik K, Strobl C, Zeileis A, Hothorn MT. Package party, 2014. Available at: https://cran.r-project.org/web/packages/party/party.pdf.
47. Therneau TM, Atkinson B, Ripley MB. The rpart package, 2010. Available at: http://ftp.uni-bayreuth.de/math/statlib/R/CRAN/doc/packages/rpart.pdf.
48. Everitt BS, Hothorn, T. A handbook of statistical analyses using R., Boca Raton: Chapman and Hall/CRC, 2006.
49. Dusseldorp E, Doove L, Van Mechelen I. Quint: An r package for identification of subgroups of clients who differ in which treatment alternative is best for them. Behav Res Methods 2016;48:650–63.
50. Carroll KM, Ball SA, Nich C, Martino S, Frankforter TL, Farentinos C, et al. Motivational interviewing to improve treatment engagement and outcome in individuals seeking treatment for substance abuse: a multisite effectiveness study. Drug Alcohol Depend 2006;81:301–12.
51. McLellan AT, Kushner H, Metzger D, Peters R, Smith I, Grissom G, Pettinati H, et al. The fifth edition of the addiction severity index. J Subst Abuse Treat 1992;9:199–213.
52. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning for detecting interaction effects in subgroups. Presented at "IFCS 2013 – International Federation of Classification Societies", Tilburg University, The Netherlands, 2013. Available at: http://eeecon.uibk.ac.at/ zeileis/papers/IFCS-2013.pdf.
53. Seibold H, Hothorn T, Zeileis A. Generalised linear model trees with global additive effects (submitted). Available on: arXiv:1612.07498 (https://arxiv.org/abs/1612.07498)

Supplemental Material: The online version of this article (DOI: 10.1515/ijb-2016-0068) offers supplementary material, available to authorized users.

Reproduced with permission of copyright owner.
Further reproduction prohibited without permission.