Research Article

Statistics in Medicine

Received 14 September 2011, Accepted 10 July 2013 Published online 6 August 2013 in Wiley Online Library

(wileyonlinelibrary.com) DOI: 10.1002/sim.5933

Qualitative interaction trees: a tool to identify qualitative treatment–subgroup interactions

Elise Dusseldorp[a,b*†] **and Iven Van Mechelen**[b]

When two alternative treatments (A and B) are available, some subgroup of patients may display a better outcome with treatment A than with B, whereas for another subgroup, the reverse may be true. If this is the case, a qualitative (i.e., disordinal) treatment–subgroup interaction is present. Such interactions imply that some subgroups of patients should be treated differently and are therefore most relevant for personalized medicine. In case of data from randomized clinical trials with many patient characteristics that could interact with treatment in a complex way, a suitable statistical approach to detect qualitative treatment–subgroup interactions is not yet available. As a way out, in the present paper, we propose a new method for this purpose, called QUalitative INteraction Trees (QUINT). QUINT results in a binary tree that subdivides the patients into terminal nodes on the basis of patient characteristics; these nodes are further assigned to one of three classes: a first for which A is better than B, a second for which B is better than A, and an optional third for which type of treatment makes no difference. Results of QUINT on simulated data showed satisfactory performance, with regard to optimization and recovery. Results of an application to real data suggested that, compared with other approaches, QUINT provided a more pronounced picture of the qualitative interactions that are present in the data. Copyright © 2013 John Wiley & Sons, Ltd.

Keywords: qualitative interaction; moderator; subgroup analysis; binary tree; partitioning; treatment efficacy

1. Introduction

1.1. Problem

When several treatment alternatives are available for a certain disease, an important question is which of these alternatives is most efficacious. The gold standard method to answer this question is a randomized controlled trial (RCT). In this paper, we focus on the situation in which two treatment alternatives are available (A and B) and a two-arm RCT has been performed. If, in the population, the mean outcome of the patients receiving A is better than the one of the patients receiving B, A is seen as more efficacious or effective than B.

Beyond the question of general efficacy or effectiveness, a frequently occurring question is whether the difference in effect between the two treatments is equal for all subgroups of patients. A subgroup or subset analysis may identify that this is not the case (Figure 1a). In formal terms, such a situation is called a treatment–subgroup interaction, subgroup–treatment effect interaction [1], or treatment–covariate interaction [2]. The patient characteristic(s) defining the subgroups then are called 'moderators of treatment effects' [3] or 'treatment effect modifiers'. We would like to emphasize that each of the subgroups in Figure 1 (subgroup a, b, c, and d) may be defined by several patient characteristics. In general, subgroup analyses identify the differential efficacy of a treatment, starting from the idea that the size of the treatment effect may differ across individuals. We use the term subgroup analysis to refer to all types of analysis involving subgroups of patients who are assigned to treatments (in line with [1]).

[a] *Statistics Group, Netherlands Organization for Applied Scientific Research (TNO), Wassenaarseweg 56, Leiden, The Netherlands*
[b] *Department of Psychology, Katholieke Universiteit Leuven, Tiensestraat 102 – bus 3713, Leuven, Belgium*
**Correspondence to: Elise Dusseldorp, Statistics Group, TNO, PO Box 2215, 2301 CE Leiden, The Netherlands.*
[†] *E-mail: elise.dusseldorp@tno.nl*

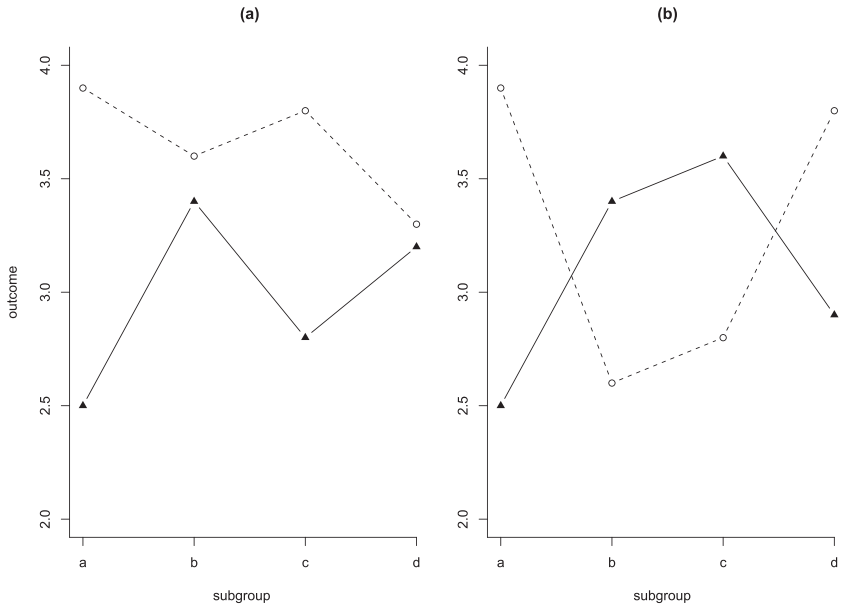

Figure 1. Two examples of treatment–subgroup interactions (▲, treatment A; ○, treatment B): (a) quantitative interaction with respect to the treatment variable and (b) qualitative interaction.

A special type of treatment–subgroup interaction occurs if for some subgroups of patients, one treatment is better than the other, whereas for other subgroups, the reverse is true. Such a situation is shown in Figure 1b. This type of interaction is called disordinal with respect to the treatment variable T [4, 5]. Such interactions, where the difference in treatment effect in one subgroup has a different sign than in another subgroup, are also referred to as qualitative interactions [2], as opposed to quantitative (or ordinal) interactions, where the difference in treatment effects has the same sign in all subgroups (Figure 1a). Qualitative interactions, unlike a number of quantitative interactions, cannot be removed by the choice of a different model [2] or by monotonic transformations of the outcome variable. An example of a qualitative interaction has been reported by Behrendt and Gehan [6] in a study on adults with acute leukemia. From this study, it appeared that experimental treatment (amsocrine plus OAP, i.e., a combination of vincristine, cytosine arabinoside, and prednisone) is superior to standard treatment (adriamycin plus OAP) in patients with unfavorable prognosis and inferior to standard treatment in patients with favorable prognosis.

In the presence of a qualitative treatment–subgroup interaction, the question 'Which treatment is better, A or B?' becomes meaningless and should be replaced by 'Which treatment is best for which kind of patients?' [4, 7]. The moderator variable(s) contributing to the qualitative interaction(s) then identify for whom and under which circumstances treatment A is better than B and for whom the reverse is true. As such, they represent important patient characteristics that may be used in the future to set up an optimal treatment assignment strategy to support healthcare decision makers [7]. It is, therefore, essential to uncover qualitative treatment–subgroup interactions with an appropriate statistical method. The development of such a method is the key challenge we want to address in the present paper.

In this development, our primary focus will be on a typical RCT context that involves a large number of potentially relevant moderator variables, without clear a priori hypotheses on the nature of the subgroups involved in qualitative treatment–subgroup interactions, and with the subgroups possibly being defined in terms of complex patterns of values of moderator variables. All this implies that the subgroups are not known in advance but are to be induced during the actual data analysis, rather. Moreover, ideally the qualitative interactions should be not only statistically but also clinically significant, with sizeable subgroup between-treatment differences of varying signs that are most relevant for clinical practice. In quite some cases, such an outcome could be facilitated by including in the induction a subgroup of patients for whom type of treatment (A or B) makes no difference (i.e., an 'indifference group', introduced before as 'region of uncertainty' [2, 8]).

1.2. Previous work

Previous methodological work on the study of treatment–subgroup interactions primarily pertained to the situation in which clear a priori hypotheses exist about which subgroups are involved in the interactions

and in which the subgroups can be defined by means of one or a small number of patient characteristics only. Most methods that have been studied at this point stem from the area of linear multivariate analysis. Important special instances include, for the case of categorical moderators only, analysis of variance (possibly in conjunction with a predefined contrast coding of the hypothesized effects to increase power [9, p. 388]); for the case of continuous or mixed continuous-categorical patient characteristics, several forms of regression analysis with suitable interaction terms have been studied, variously referred to as moderated regression analysis [10] and aptitude treatment interaction analysis [11]. To capture more complex and nonlinear interaction effects, such approaches have been extended in several directions, making use of, for example, fractional polynomials [12], random effects [13], and generalized additive modeling techniques [14].

For the context that constitutes the focus of the present paper, with a large number of potential moderators and absence of clear a priori hypotheses, several authors have warned against multiplicity problems and spurious interactions that cannot be replicated in follow-up studies [1, 15]. Keeping in mind these warnings, so far, a few interesting methods have been developed. These methods induce subgroups involved in a treatment–subgroup interaction from the data. All methods in question are of a recursive partitioning type. They are the regression trunk model implemented as Simultaneous Threshold Interaction Modeling Algorithm (STIMA) [16, 17], Interaction Trees [18, 19], Virtual Twins [20], and Subgroup Identification Based on Differential Effect Search (SIDES) [21].

The goal of STIMA and Interaction Trees is to partition the total group of patients into subgroups that differ as much as possible in relative treatment effectiveness; this implies that the two methods look for subgroups involved in an as large as possible treatment–subgroup interaction. The other two methods, Virtual Twins and SIDES, start by considering one of the two treatment alternatives as the reference treatment and the other as the alternative treatment; subsequently, the methods aim at identifying specific subgroups of patients in which the alternative treatment outperforms as much as possible the reference treatment, while disregarding all other patients in the sample (for a detailed comparison of the four methods, see [22]). As the goal of STIMA and Interaction Trees is related more closely to the one of the present paper, from now on, we will primarily focus on these two methods.

STIMA is an elaboration of a an earlier developed method for the detection of treatment–subgroup interactions [16]. It is based on a hybrid methodology, combining a multiple regression model and a tree from which interaction terms in the regression model are derived. Within the context of treatment–subgroup interactions, the dependent variable in the regression model is treatment outcome. STIMA starts with a regression model that contains main effects of treatment type and of all moderators, next to a tree for which the first split is made on the basis of treatment variable T. In the remainder of the algorithmic process, the tree will undergo a sequential splitting, with each split being based on one of the moderators and with each node of the tree implying a treatment–moderator interaction term that is added to the regression model. In each step of the algorithm, STIMA will search among all leaves of the current tree, among all moderators, and among all split points for the split that is associated with the interaction term that induces the highest increase in variance accounted for by the regression model.

Interaction Trees aim at accounting for heterogeneity in differential treatment effectiveness in two-arm RCTs by a sequential tree building process. Like for STIMA, this process is regression model based yet only making use of local regression models. That is to say, given a leaf of the current tree, a moderator, and a split point, two local regression models (with treatment outcome as dependent variable) are compared: (1) a regression model that includes main effects of treatment type T and of an indicator variable D for a split on the basis of the moderator and split point under study and (2) a regression model that includes the same two main effects *plus* an interaction term, that is, the product between T and D. For the leaf under study, the split (i.e., combination of a moderator and a split point) is selected for which model (2) implies the most significant gain over model (1).

Both STIMA and Interaction Trees allow the user to identify subgroups of patients that are involved in treatment–subgroup interactions from data sets with many patient characteristics that could possibly interact with treatment. Moreover, both methods also involve pruning procedures to avoid solutions with spurious interaction effects. Yet, a shortcoming of the two methods is that they address treatment–subgroup interactions in general and that the user has no control over the type of interactions involved in the tree. This may be a significant drawback if especially qualitative interactions are of interest. For example, in the presence of strong quantitative interaction effects, qualitative interactions may remain undetected. An additional shortcoming is that in the splitting processes of STIMA and Interaction Trees, indifference groups are not considered, which may hamper the identification of clinically meaningful qualitative interactions.

1.3. Aim of the present paper

The present paper proposes a novel sequential partitioning method exactly designed to remedy for the shortcomings mentioned earlier. This method is called QUalitative INteraction Trees (QUINT). Making use of a set of novel partitioning criteria, the method first performs a check if qualitative treatment–subgroup interactions are present in the data. If this is the case, QUINT automatically identifies the combinations of dichotomized moderators that are most important for qualitative treatment–subgroup interactions. The result of a QUINT analysis is a binary tree that implies a partitioning of the total sample in three groups of patients: those for whom treatment A is better than treatment B, those for whom B is better than A, and those for whom it does not make any difference.

The remainder of this paper is organized as follows. First, we describe the conceptual basis of QUINT, along with the associated algorithm (Section 2). Second, we test the performance of the method with a simulation study, in which we will also pay special attention to the risk of inferential errors (Section 3). Third, we apply QUINT to the data set of breast cancer patients from Scheier *et al.* [23] (Section 4); in this application, we will also compare the performance of QUINT with that of STIMA and Interaction Trees. Discussion points and concluding remarks are given in the final section.

2. Method

2.1. Overview

We start from a group of N patients randomly assigned to one of two treatments A and B. All patients are measured before and after the treatment. Before the treatment, a group of categorical and/or continuous background characteristics of the patients is measured (e.g., severity of disease), that is, a set of baseline variables. After the treatment, one primary continuous outcome variable is measured. The outcome variable can also be measured twice: before and after the treatment. In that case, the change score, or the slope of response over time, or time to an event can be used as outcome for QUINT. The goal of QUINT is to find the best partition of the total group of patients on the basis of the baseline variables into two or three mutually exclusive and exhaustive subgroups (i.e., partition classes) that are characterized as follows: In the first (nonempty) subgroup ($\wp_1$), the patients assigned to treatment A show a clearly better outcome than the patients assigned to B; in the second (nonempty) subgroup ($\wp_2$), the reverse is true; and in the third (optional) subgroup ($\wp_3$), the patients assigned to A show more or less the same outcome as the patients assigned to B. It is important to note that the subgroups may comprise one or several types of patients as defined by different (combinations of) patient characteristics.

We are looking for a partition that is optimal in the sense that the qualitative treatment–subgroup interaction has the largest possible practical significance. This means that the interaction should imply the presence of a sizeable subgroup of patients for which assignment to treatment A would be strongly preferable, as well as the presence of another sizeable subgroup of patients for which assignment to treatment B would be strongly preferable. To achieve this, two conditions with regard to the subgroups $\wp_1$ and $\wp_2$ need to be satisfied, each of which implies two subconditions that are to be met simultaneously: (a) In both $\wp_1$ and $\wp_2$, the difference in outcome between treatments A and B should be as large as possible and (b) both $\wp_1$ and $\wp_2$ should comprise as many patients as possible. We will further call the first condition the 'Difference in treatment outcome component' and the second one the 'Cardinality component'. In Section 2.3, we will propose exact measures for these components: with regard to Difference in treatment outcome component, a measure for the extent to which one treatment is preferred to another treatment in both $\wp_1$ and $\wp_2$, and with regard to the Cardinality component, a measure for the extent to which the subgroups assigned to $\wp_1$ and $\wp_2$ are sizeable. Finally, for the qualitative treatment–subgroup interaction to have really practical significance, both component conditions need to be satisfied. Hence, our partitioning criterion will imply that the conjunction of the two components will be *maximized*.

We focus on partitions that can be obtained through a binary tree based on dichotomizations of the patient characteristics. As examples, one may consider the trees as represented in Figure 2. The QUINT algorithm starts with a tree consisting of a single node, that is, the root node containing all patients. Next, it follows a stepwise binary splitting procedure. This procedure implies that in each step, a node, a baseline characteristic, a split of that characteristic, and an assignment of the terminal nodes or leaves of the current tree to ($\wp_1$, $\wp_2$, $\wp_3$) are chosen that maximize the partitioning criterion. Note that this means that after *each* split, *all* leaves of the tree are re-assigned afresh to the subgroups $\wp_1$, $\wp_2$, $\wp_3$; hence, the QUINT procedure is nonrecursive. In the following subsections, we formalize all the above.

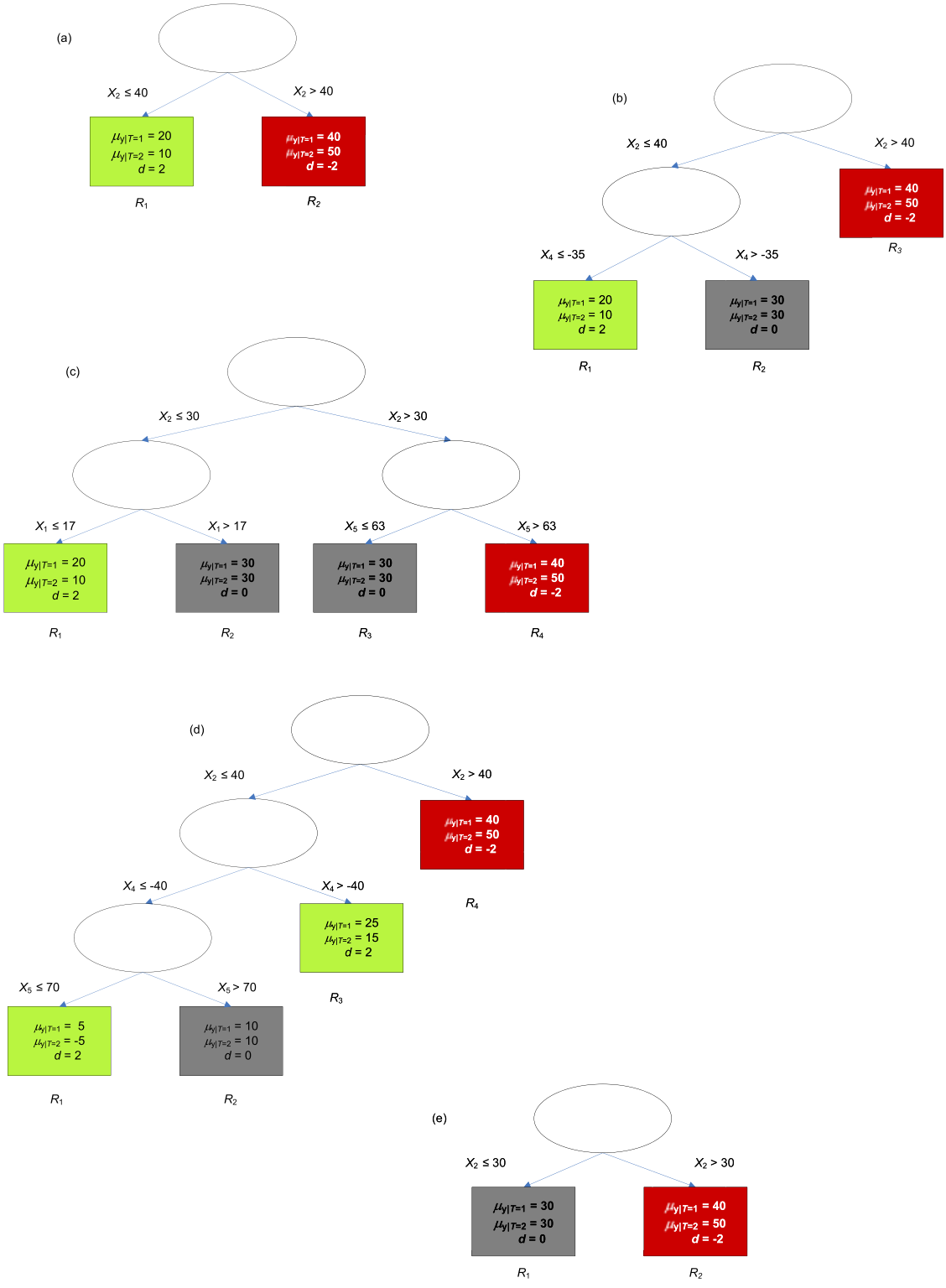

Figure 2. Simulated data sets were generated from five true tree models: (a) model A; (b) model B; (c) model C; (d) model D; and (e) model E. Each leaf (rectangle) contains the conditional outcome means for the two treatment groups ($\mu_{Y|T=1}$ and $\mu_{Y|T=2}$) and Cohen's effect size (d), expressed as the standardized mean difference between $T = 1$ and $T = 2$. Assignment of the leaves to the three partition classes is represented in green, $\wp_1$; red, $\wp_2$; and dark grey, $\wp_3$.

2.2. Concepts and notation

Let variable T be a treatment indicator variable, with levels 1 and 2 referring to treatments A and B, respectively. We further denote the baseline characteristics by $X_1, \ldots, X_J$ and the outcome variable by Y.

We look for a tripartition $\{\wp_1, \wp_2, \wp_3\}$ of the value range of $(X_1, \ldots, X_J)$ for which it largely holds that

$$\text{if } (X_1, \ldots, X_J) \in \wp_1: \quad Y|T=1 > Y|T=2,$$

$$\text{if } (X_1, \ldots, X_J) \in \wp_2: \quad Y|T=2 > Y|T=1,$$

and

$$\text{if } (X_1, \ldots, X_J) \in \wp_3: \quad Y|T=1 \approx Y|T=2,$$

where '>' means 'preferable to' and '≈' means 'approximately equal to' in terms of raw scores on Y or of effect sizes (also see Section 2.3.1). The tripartitions are based on a binary tree. We denote a leaf (i.e., region) of such a tree by R_ℓ, with $\ell = 1, \ldots, L$ and with L being the total number of leaves. For example, the set of leaves of the tree of Figure 2c is $\{R_1, R_2, R_3, R_4\}$, where the nodes have been numbered from left to right.[‡] Patients belong to a particular node (i.e., a patient type) on the basis of their scores on the baseline characteristics.

To link a binary tree to a tripartition $\{\wp_1, \wp_2, \wp_3\}$, we define an assignment function f that assigns each node $\{R_1, \ldots, R_L\}$ to one of the partition classes $\{\wp_1, \wp_2, \wp_3\}$. For example, if $f(2) = 3$, this means that node R_2 is assigned to $\wp_3$. In Figure 2c, it holds that $f(1) = 1$; $f(2) = 3$; $f(3) = 3$; and $f(4) = 2$. It then follows that $\wp_1 = R_1$, $\wp_2 = R_4$, and $\wp_3 = R_2 \cup R_3$.

2.3. Partitioning criterion

As mentioned before, our partitioning criterion includes two components that are optimized simultaneously, the Difference in treatment outcome component and the Cardinality component. In the following, we will discuss both components successively; next, we will discuss their combination into a single overall optimization criterion.

2.3.1. Difference in treatment outcome component. We consider two possible specifications of the Difference in treatment outcome component on the level of the leaves of the tree: (a) the difference in means of outcome Y and (b) the corresponding effect size, quantified through Cohen's d [24]. Both specifications imply that the difference in treatment outcome for a node R_ℓ is defined as

$$\alpha_\ell \left(\bar{Y}_{T=1,\ell} - \bar{Y}_{T=2,\ell}\right). \tag{1}$$

If $\alpha_\ell = 1$, the expression in (1) denotes a difference in treatment means; if $\alpha_\ell = 1/s_\ell$, it denotes the effect size Cohen's d [24, pp. 66–67], with s_ℓ being defined as

$$s_\ell = \sqrt{\frac{(n_1 - 1)s^2_{T=1,\ell} + (n_2 - 1)s^2_{T=2,\ell}}{n_{T=1} + n_{T=2} - 2}}, \tag{2}$$

where n_1 and n_2 denote the sample sizes of the treatment groups (respectively, $T = 1$ and $T = 2$). In other words, s_ℓ is the pooled within-sample estimate of the population standard deviation of the treatment groups in leaf R_ℓ. As a variant of s_ℓ, one could consider the average standard deviation in the two treatment groups, defined as

$$s'_\ell = \sqrt{\frac{s^2_{T=1,\ell} + s^2_{T=2,\ell}}{2}}. \tag{3}$$

In numerical work, we found similar solutions using s'_ℓ and s_ℓ. A plausible reason for this is that Cohen's d is used here purely as a descriptive statistic of the magnitude of the treatment effect in the sample, with

[‡] *After a split, the nodes that form the leaves of a tree at that moment are renumbered from left to right.*

no inferences being made about population values. A further discussion of the choice of the difference in treatment outcome (i.e., the value of α_ℓ) is given in Supplementary materials A.1.[§]

The overall difference in treatment outcome of the nodes assigned to $\wp_1$ (denoted by D_1) can be computed as a weighted average of the difference in treatment outcome across all the nodes R_ℓ:

$$D_1 = \frac{\sum_{\ell=1}^{L} I(f(\ell)=1)\,\#R_\ell\;\alpha_\ell\left(\bar{Y}_{T=1,\ell} - \bar{Y}_{T=2,\ell}\right)}{\sum_{\ell=1}^{L} I(f(\ell)=1)\,\#R_\ell}, \tag{4}$$

where $\#R_\ell$ denotes the number of patients in node R_ℓ, and $I(f(\ell)=1)$ denotes an indicator function of the leaves assigned to $\wp_1$. Similarly, the overall difference in treatment outcome of the leaves R_ℓ assigned to $\wp_2$ is expressed as

$$D_2 = \frac{\sum_{\ell=1}^{L} I(f(\ell)=2)\,\#R_\ell\;\alpha_\ell\left(\bar{Y}_{T=2,\ell} - \bar{Y}_{T=1,\ell}\right)}{\sum_{\ell=1}^{L} I(f(\ell)=2)\,\#R_\ell}. \tag{5}$$

One may note that (4) and (5) imply that in calculating an overall measure of difference in treatment outcome for a partition class, more weight is given to nodes within that class where the difference can be more reliably estimated because of a larger sample size.

As explained in Section 2.1, for a qualitative treatment–subgroup interaction to be clinically or practically significant, it is of utmost importance that the difference in treatment outcome is sizeable in *both* $\wp_1$ and $\wp_2$. The Difference in treatment outcome component is therefore put equal to the *product* $D_1 \times D_2$, with D_1 and D_2 as given in (4) and (5).

2.3.2. Cardinality component. Regarding the Cardinality component, we first calculate for each of the two partition classes $\wp_c$ ($c = 1, 2$) the sum of the cardinalities of the nodes assigned to that partition class, which may be expressed as

$$\left(\sum_{\ell=c}^{L} I(f(\ell)=c)\,\#R_\ell\right).$$

Similarly to the difference in treatment outcome and as explained in Section 2.1, for a qualitative treatment–subgroup interaction to be practically significant, it is of utmost importance that the cardinality of *both* $\wp_1$ and $\wp_2$ is sizeable. The Cardinality component is therefore put equal to the *product* of the cardinalities of the two partition classes.

2.3.3. Combination of the two components into a single overall partitioning criterion. Practical significance requires that the Difference in treatment outcome component and the Cardinality component are to be maximized simultaneously (Section 2.1). In principle, an overall partitioning criterion then could be the product of the two components in question. From a practical point of view, however, it is more convenient to put the two components on a log-scale, such that the final partitioning criterion can be expressed in an additive rather than a multiplicative way.

Before being able to formulate this final criterion, two obstacles still have to be removed. First, because D_1 and D_2 can take values lower than 1, their logarithms may become negative. To remedy for this, we take the logarithm of $1 + D_1$ (respectively $1 + D_2$). Second, to properly combine the Difference in treatment outcome and Cardinality components, the two components need to be put on comparable measurement scales. For this purpose, we give the two components suitable, well-defined weights, which are derived by setting the realistic maximum of both components after weighting about equal (details of the

[§] *Supporting information may be found in the online version of this article.*

derivation and the resulting values of the weights are given in Supplementary materials A.2). Taking into account all the aspects mentioned earlier, our global partitioning criterion (C) then reads as follows:

$$C = w_1 \left[\log(1 + D_1) + \log(1 + D_2)\right] + w_2 \left[\log\left(\sum_{\ell=1}^{L} I(f(\ell) = 1)\, \#R_\ell\right) + \log\left(\sum_{\ell=1}^{L} I(f(\ell) = 2)\, \#R_\ell\right)\right], \quad (6)$$

with D_1 and D_2 being defined in (4) and (5), respectively, and with w_1 denoting the weight of the Difference in treatment outcome component and w_2 that of the Cardinality component.

As an aside, one may note that the maximization of the Cardinality component forces the cardinalities of $\wp_1$ and $\wp_2$ to be approximately equal, which implies that D_1 and D_2 as included in the Difference in treatment outcome component are calculated on the basis of comparable sample sizes.

2.4. *The sequential partitioning algorithm*

2.4.1. Stepwise procedure. QUINT uses a stepwise binary tree algorithm that maximizes partitioning criterion C. The algorithm starts from a tree consisting of a single node containing all patients. During the splitting process, the algorithm takes into account several stopping criteria. These criteria determine whether a solution is admissible or not, with only admissible solutions being considered in the optimization procedure. We describe first the stepwise procedure and then the stopping criteria.

In each step of the algorithm, all leaves of the tree as obtained from the previous step are considered as candidate parent nodes. Two substeps are then performed. In the first substep, the algorithm looks in each candidate parent node for the optimal split in terms of an optimal combination of three ingredients (a triplet): a splitting variable, a split point, and an admissible assignment of all the leaves of the tree *after* the split. For this purpose, each baseline characteristic acts as a candidate splitting variable. For each of these candidates, first from the total set of observed values, the subset of admissible split points is determined; second, the split point and admissible assignment are chosen that induce the highest value of the partitioning criterion C. This substep then is concluded by selecting across all candidate splitting variables the triplet that implies the highest value of C; this is retained as the optimal triplet for the candidate parent node under study.

In the second substep, the values of C associated with the optimal triplets are compared across all candidate parent nodes, and the node with the highest value is chosen. If this value is higher than that of the tree resulting from the previous step, the chosen node then is split into two child nodes (on the basis of the characteristic and split point as included in the optimal triplet associated with that node).

2.4.2. Stopping criteria. The QUINT algorithm stops when a split can no longer be found that implies a higher value of C. The user, however, may stop the algorithm earlier, by specifying a priori the maximum possible number of leaves ($L_{\text{upperlimit}}$). The QUINT algorithm further takes into account four additional stopping criteria, which can be regarded as boundary conditions.

The first criterion is checked after the first split only, when the tree has two leaves; these leaves have to be assigned to $\wp_1$ and $\wp_2$ (due to the *nonempty partition class condition*, which will be explained later). The criterion then reads that the absolute value of the standardized mean difference in treatment outcome in each of the two leaves should exceed a critical minimum value ($d_{\min}$). This can be seen as a check of whether a qualitative interaction is present in the data and therefore is called the *qualitative interaction condition*. If this condition is met, a tree is grown; otherwise, no tree is grown.

The second criterion pertains to the tree growing process. It reads that in each leaf, a minimum number of patients is in treatment A and a minimum number in treatment B. For reliable estimation of each treatment mean, the number of patients is controlled per treatment. This criterion is referred to as the *minimal sample size per treatment condition*.

The final two criteria pertain to the assignment of all leaves of the three to the partition classes after each split. The first of them reads that the partition classes $\wp_1$ and $\wp_2$ are not empty and will be referred to as the nonempty partition class condition. The second reads that a node can be assigned to $\wp_1$ only if in that node, the mean outcome of the patients in treatment A exceeds that of the patients in treatment B; similarly, a node can be assigned to $\wp_2$, only if the mean outcome of the patients in A is lower than that of the patients in B. This criterion will further be referred to as the *mean difference per node condition*.

It may happen that, for a specific data set, in the first step of the QUINT procedure, one or more of the four criteria outlined earlier cannot be met, and hence, no admissible solutions for this data set are

available. For example, if for every partition of the patients on the basis of splits on threshold values of the baseline variables, it holds that in each node, the mean outcome of the patients in A exceeds that of the patients in B, one will never be able to find a qualitative treatment–subgroup interaction.

One may finally note that the specification of the QUINT criterion and stopping criteria imply several user-defined choices. An overview of these choices, along with recommendations and default values, can be found in Supplementary materials A.

2.5. Pruning

The tree growing process of QUINT stops if no more parent node can be found with an admissible triplet and a higher value of C than in the previous step (or if the total number of leaves equals $L_{\text{upperlimit}}$). This may result in a large tree that fits the data at hand perfectly but that will not fit future data. To avoid this so-called overfitting and to increase the predictive validity of the final tree, a generally accepted strategy within the domain of tree-based methods is to prune the maximal tree back to some optimal subtree.

For the pruning of QUINT, we appeal to a bias-corrected bootstrap procedure as proposed by Efron [25] and as applied to tree-based models by LeBlanc and Crowley [26]. Our pruning procedure further implies a simplified form of cost complexity pruning [26,27] and relies on the fact that the nonrecursive stepwise algorithm automatically yields a sequence of nested subtrees, for which the number of leaves (L) may act as a complexity parameter. A detailed description of the pruning procedure in QUINT can be found in Supplementary materials B.

3. Simulation study

3.1. Motivation

In this section, we want to evaluate the optimization and recovery performance of QUINT. Optimization performance pertains to whether the QUINT algorithm is successful in identifying a solution with optimal value for criterion C, given in (6). This question is linked to the sequential nature of the QUINT algorithm: The stepwise partitioning procedure guarantees that within each split, an optimal triplet is found but not that the solution after several splits is still globally optimal. We would like to get an idea to what extent this is the case.

Recovery performance from its part pertains to the extent to which the QUINT algorithm is successful in retrieving the true structure underlying the data. At this point, four structural aspects (RP1–RP4) can be distinguished. The first three of these pertain to the structure of the true underlying tree. In particular, RP1 relates to the presence/absence of a qualitative treatment–subgroup interaction in this tree. That is, we want to know the probability that QUINT decides wrongly that a qualitative interaction is present in data generated from a true tree structure *without* a qualitative treatment–subgroup interaction (type I error [RP1a]); also, we would like to know to which extent QUINT decides wrongly that a qualitative interaction is *not* present in data generated from a true tree structure *with* a qualitative treatment–subgroup interaction (type II error [RP1b]). In other words, the type I error rate represents the probability that the QUINT solution identifies spurious qualitative interaction effects and the type II error rate represents the probability that QUINT fails in detecting true qualitative interaction effects.

Regarding RP2, one may wish to know whether, given an underlying true tree with a qualitative treatment–subgroup interaction that QUINT has correctly detected, QUINT is also successful in identifying the complexity of the true tree; or, stated in other words, one may wish to know the quality of performance of the pruning rule. Furthermore, given the same situation and given the true complexity of the underlying tree, one may wish to know to what extent QUINT is able to recover the structure of the true tree in terms of the true splitting variables and the true split points (RP3). A fourth and final aspect of recovery (RP4) pertains to the assignment of the observations to the three partition classes.

3.2. Design

Artificial data sets were generated that differed in the complexity of the binary tree as involved in the treatment–subgroup interaction. Each data set consisted of N observations on one binary treatment variable T, J continuous baseline variables or covariates ($X_1, \ldots, X_J$), and one continuous outcome Y. Variables X_1 to X_J were multivariate normally distributed, with the value of μ_{X_j} for the splitting variables used in the true models (X_1, X_2, X_4, X_5; Figure 2) being fixed at 10, 30, −40, and 70, respectively, with the value μ_{X_j} for the other variables being drawn from a discrete uniform distribution on the

interval $[-70, 70]$, and with for all variables $\sigma_{X_j} = 10$. The correlation between the covariates was a design factor (see following text). T was Bernoulli distributed, with $\theta = 0.50$, and independent from $X_1, \ldots, X_J$. The distribution of Y was based on five different true binary tree structures: models A to E (Figure 2a–e). Models A through D all contain one or more qualitative interactions, while Model E does not contain a qualitative interaction but a quantitative one; this latter model is used to investigate the type I error. Per leaf and per treatment condition, Y was normally distributed, with $\sigma_Y = 5$ and with μ_Y depending on the particular combination of the leaf and treatment condition in question. We fully crossed four design factors: (1) the *sample size N* (having five levels: 200, 300, 400, 500, and 1000); (2) the total *number of covariates J* (having three levels: 5, 10, or 20); (3) the value for the *difference in treatment outcome* in the leaves assigned to $\wp_1$ or $\wp_2$ ($\mu_{Y|T=1} - \mu_{Y|T=2}$ was set at 2.5, 5.0, and 10.0, respectively, for the leaves assigned to $\wp_1$, and $\mu_{Y|T=1} - \mu_{Y|T=2}$ was set at -2.5, -5.0, and -10.0, respectively, for the leaves assigned to $\wp_2$, implying effect sizes [d] in these leaves with a medium value [$d = 0.50$ or -0.50], moderately large [$d = 1$ or -1], and very large [$d = 2$ or -2] value); and (4) the *intercorrelation* between the covariates (having two levels: $\rho_{j\,j'} = 0.20$ and $\rho_{j\,j'} = 0.20$, $\forall\, j \neq j' = 1, \ldots, J$). For each cell of the design, 100 data sets were generated. This resulted in $5 * 3 * 3 * 2 * 100 = 9000$ data sets for each of the five models.

3.3. Analysis

QUINT was applied to each data set using the treatment effect size criterion (Section 2.3.1). Thus, in total, 45,000 QUINT analyses were performed. The maximum tree size ($L_{\text{upperlimit}}$) was fixed a priori at 8, and default values were used for the other stopping criteria, the weights, and the number of bootstrap samples (given in Supplementary materials A and B). The 1-SE pruning rule was applied to determine the optimal size of the tree. The analyses were performed in R [28] with the R package 'quint', which is available from the authors upon request.

3.4. Results

3.4.1. Optimization performance. Regarding optimization performance, we wanted to know whether QUINT yields globally optimal solutions. Yet, the globally optimal solution of the QUINT problem is in general unknown. As a way out, as a lower bound for the globally optimal criterion values, we used the criterion values (C) of the true structure that was used to generate the data. We then compared these with the values of the QUINT solutions ($\hat{C}$), for each split of the tree (up to the true number of splits for each data set). A nonsuspected solution was defined as a solution with $\hat{C} - C > -0.05$, with the value of 0.05 being chosen by taking into account numerical precision. The proportion of nonsuspected solutions was determined for all true models that included a qualitative interaction (i.e., models A through D). The results showed that across all models, design factors, and splits, the mean value of the proportion of nonsuspected solutions was high (0.97), indicating an overall good optimization performance. The mean value decreased with increasing size of the tree: After the first, second, and third splits, the mean proportion was 1.00, 0.97, and 0.91, respectively.

3.4.2. Recovery performance.

(RP1a) *Probability of type I errors.* With regard to the type I error, remember that the QUINT algorithm decides that a qualitative interaction is present, unless the qualitative interaction condition is violated. This condition states that, after the first split, the minimum of the absolute value of the effect sizes in the two leaves should exceed the value of $d_{\min}$. The type I error of QUINT was evaluated making use of the model E data sets (Figure 2) for various values of $d_{\min}$ (between 0.20 and 0.40). We then counted for each cell of our design and for various values of $d_{\min}$ the proportion of QUINT solutions without a violation of the qualitative interaction condition (i.e., the proportion of solutions for which QUINT wrongly decided that a qualitative interaction was present, or the type I error rate). The resulting proportions were subjected to an analysis of variance (ANOVA) with the four design factors and $d_{\min}$ as independent variables and with the five-way interaction being used as error term. The ANOVA results revealed that type I error rate was influenced mainly (partial $\eta^2 > 0.80$) by the value of $d_{\min}$, *sample size*, *difference in treatment outcome*, and the two-way interactions $d_{\min}$ * *sample size*, and $d_{\min}$ * *difference in treatment outcome*. Error rates averaged across the levels of the two less important factors, *number of covariates* and *intercorrelation*, are presented in Table I. For the smaller sample sizes ($N = 200$ and

Table I. Proportion of solutions for data from a true model with no qualitative interaction (model E, Figure 2) where the QUINT algorithm wrongly decided that a qualitative interaction is present (type I error rate) for various values of the parameter of the qualitative interaction condition (d_{min}).

		d_{min}				
DT	*N*	0.20	0.25	0.30	0.35	0.40
M	200	0.72	0.61	0.47	0.32	0.20
	300	0.51	0.37	0.24	0.14	0.06
	400	0.45	0.29	0.13	0.06	0.01
	500	0.29	0.16	0.08	0.02	0.00
	1000	0.10	0.03	0.01	0.00	0.00
L	200	0.56	0.47	0.37	0.28	0.23
	300	0.39	0.30	0.22	0.14	0.08
	400	0.29	0.18	0.11	0.07	0.03
	500	0.21	0.12	0.06	0.02	0.00
	1000	0.07	0.03	0.01	0.00	0.00
XL	200	0.39	0.31	0.22	0.18	0.14
	300	0.27	0.18	0.12	0.07	0.04
	400	0.18	0.11	0.06	0.04	0.01
	500	0.13	0.07	0.03	0.01	0.00
	1000	0.03	0.02	0.01	0.00	0.00

Results are presented separately for simulated data with varying true size of the difference in treatment outcome (DT) (i.e., medium [M], moderately large [L], and very large [XL]) and with varying sample size N. Results have been averaged across the levels of the factors number of covariates (5, 10, or 20) and intercorrelation between covariates ($\rho = 0$ or $\rho = 0.20$).

300), the type I error rate appeared to be high (>0.15) for most values of d_{min}. For the other sample sizes and for d_{min} values of 0.30 and higher, the type I error seemed largely acceptable.

(RP1b) *Probability of type II errors.* For the data sets generated from the models with a true qualitative interaction (i.e., models A through D), we calculated for each cell of our design and for various values of d_{min} the proportion of solutions *with* a violation of the qualitative interaction condition (i.e., the proportion of solutions for which QUINT wrongly decided that a qualitative interaction was *not* present, or the type II error rate). The ANOVA results revealed that type II error rate was influenced mainly (partial $\eta^2 > 0.80$) by the value of d_{min}, *sample size*, *difference in treatment outcome*, and the three-way interaction d_{min} * *sample size* * *difference in treatment outcome*. Error rates averaged across the levels of the two less important factors are presented in Table II. This table shows that the effect of the *difference in treatment outcome* was most important: For a medium-sized difference in treatment outcome, the type II error appeared to be acceptable (<0.20), only for the simple model (model A) when d_{min} values were 0.30 or lower and sample sizes (N) were 400 or higher; for a moderately large difference, the type II error was acceptable (<0.20) for the less complex models (models A and B) when d_{min} values were 0.30 or lower, and for the more complex models (models C and D) when d_{min} was equal to 0.20; for a very large difference in treatment outcome, the type II error was completely negligible (<0.05).

(RP2) *Recovery of tree complexity.* For each data set, we compared the size of the pruned QUINT tree with the true tree size in terms of exact match and exact match plus or minus one. To keep a good balance between types I and II error, we used a value of $d_{min} = 0.30$ for the qualitative interaction condition. The ANOVA results revealed that the recovery rate of the true tree size was influenced mainly by *sample size*, *difference in treatment outcome*, and the two-way interaction between them. For a medium-sized and moderately large difference in treatment outcome, the recovery rate (exact match plus or minus one) decreased with increasing complexity of the tree (models C and D) in conjunction with smaller sample sizes (Table III). For a very large difference in treatment outcome, the recovery rate was satisfactory (>0.80) for almost all models (Table III).

(RP3) *Recovery of splitting variables and split points.* The ANOVA results revealed that the recovery of the splitting variables and split points was influenced by all design factors (and the three-way

Table II. Type II error: proportion of solutions for data from models A through D (Figure 2) that are wrongly indicated as without a qualitative interaction; the cells display the type II error rate for various values of $d_{\min}$ (0.20, 0.30, and 0.40).

		Model A $d_{\min}$			Model B $d_{\min}$			Model C $d_{\min}$			Model D $d_{\min}$		
DT	*N*	0.20	0.30	0.40	0.20	0.30	0.40	0.20	0.30	0.40	0.20	0.30	0.40
M	200	0.13	0.24	0.47	0.26	0.48	0.78	0.24	0.52	0.82	0.30	0.60	0.87
	300	0.11	0.21	0.39	0.30	0.60	0.87	0.30	0.67	0.94	0.40	0.76	0.96
	400	0.08	0.17	0.39	0.32	0.67	0.92	0.39	0.80	0.96	0.52	0.89	0.99
	500	0.05	0.13	0.33	0.36	0.75	0.96	0.35	0.81	0.98	0.61	0.93	1.00
	1000	0.03	0.08	0.26	0.26	0.81	0.99	0.45	0.93	1.00	0.68	0.97	1.00
L	200	0	0.01	0.04	0.05	0.14	0.33	0.11	0.24	0.50	0.18	0.42	0.68
	300	0	0.01	0.02	0.02	0.12	0.32	0.06	0.20	0.48	0.13	0.42	0.73
	400	0	0	0	0.01	0.08	0.31	0.06	0.21	0.52	0.11	0.42	0.78
	500	0	0	0	0.01	0.07	0.32	0.04	0.14	0.42	0.11	0.42	0.81
	1000	0	0	0	0	0.01	0.24	0	0.06	0.36	0.02	0.33	0.82
XL	200	0	0	0	0	0	0	0.01	0.01	0.02	0	0	0.01
	300	0	0	0	0	0	0	0	0	0.01	0	0	0
	400	0	0	0	0	0	0	0	0	0	0	0	0
	500	0	0	0	0	0	0	0	0	0	0	0	0
	1000	0	0	0	0	0	0	0	0	0	0	0	0

Results are presented separately for simulated data with varying true size of the difference in treatment outcome (DT) (i.e., medium [M], moderately large [L], and very large [XL]) and with varying sample size N. Results have been averaged across the levels of the factors number of covariates (5, 10, or 20) and intercorrelation between covariates ($\rho = 0$, or $\rho = 0.20$).

Table III. Goodness-of-recovery performance of the pruning rule of QUINT; cells display the conditional proportions of solutions with the correct true tree size (left) and the correct true tree size plus or minus one (right).

		True size Model				True size ±1 Model			
DT	*N*	A	B	C	D	A	B	C	D
M	200	0.67	0.12	0.05	0.04	0.75	0.61	0.26	0.27
	300	0.79	0.21	0.05	0.05	0.85	0.72	0.27	0.29
	400	0.86	0.16	0.06	0.07	0.92	0.81	0.31	0.17
	500	0.89	0.28	0.08	0.09	0.94	0.88	0.32	0.27
	1000	0.99	0.37	0.21	0.21	1.00	0.98	0.62	0.39
L	200	0.97	0.22	0.09	0.10	0.98	0.77	0.34	0.33
	300	0.98	0.57	0.18	0.18	1.00	0.92	0.53	0.51
	400	0.99	0.75	0.32	0.30	1.00	0.95	0.63	0.75
	500	1.00	0.86	0.55	0.38	1.00	0.96	0.77	0.89
	1000	1.00	0.95	0.93	0.60	1.00	1.00	0.99	0.98
XL	200	0.99	0.81	0.40	0.20	0.99	0.91	0.60	0.86
	300	1.00	0.96	0.75	0.30	1.00	0.98	0.84	0.93
	400	1.00	0.98	0.93	0.36	1.00	0.99	0.96	0.94
	500	1.00	0.96	0.96	0.54	1.00	0.99	0.99	0.97
	1000	1.00	0.97	0.99	0.82	1.00	1.00	1.00	0.99

Results are presented separately for simulated data with varying true size of the difference in treatment outcome (DT) (i.e., medium [M], moderately large [L], and very large [XL]) and with varying sample size N. Results have been averaged across the levels of the factors number of covariates (5, 10, or 20) and intercorrelation between covariates ($\rho = 0$ or $\rho = 0.20$).

interaction between them) except for the *intercorrelation between the covariates*. The results are given in the Supplementary Materials Table 1. Especially the combination of a medium-sized difference in treatment outcome with a smaller sample size implied a poorer recovery, with this phenomenon strengthened in case of a larger number of covariates.

(RP4) *Recovery of the assignments of the observations to the partition classes.* For each data set, the agreement (Cohen's κ) between the assignment to the partition classes of the QUINT solution and the true assignment was computed. Then, for each model, the 100 Cohen's κ values for each cell of the simulation design were collected into one matrix (resulting in matrix of 9000 by 5). Next, an ANOVA analysis was performed, with the four design factors as independent variables and the value of Cohen's κ as dependent variable.

For models A through D, the mean Cohen's κ values across all design factors varied between almost perfect to moderate: 0.93, 0.72, 0.64, and 0.56, respectively. The ANOVA results revealed that the value of Cohen's κ was influenced mainly by *sample size* and *difference in treatment outcome* and, to some extent, by the two-way interaction between them. The results are given in the Supplementary materials Table 2. Especially the combination of a medium-sized difference in treatment outcome with a smaller sample size implied a poorer recovery.

3.5. Discussion

The results of the simulation study showed that the optimization performance was satisfactory; only few solutions were found that were clearly local optima. For the first split, this finding is tautological; for further splits, this result has information value. Furthermore, the results of the types I and II error revealed that a good balance between the two can be obtained if a value of $d_{\min} = 0.30$ is used in the qualitative interaction condition. However, caution needs to be taken when sample sizes are small ($N \leqslant 300$). It should be noted that the choice of $d_{\min}$ can also be made a priori on a theoretical or clinical basis, taking into account, for example, the nature of the disorder being treated.

The type I errors found in our simulation study were comparable with those of Interaction Trees [18, Table 2, p. 149] and somewhat higher than those of STIMA [17, Table 1, p. 521]. The type II errors of QUINT were somewhat higher than those of Interaction Trees [18, Table 2, p. 149].

The evaluation of the recovery performance of the other aspects (tree complexity, tree structure, and assignment to partition classes) revealed that the results were satisfactory for all models when sample sizes were larger ($N \geqslant 400$) and true differences in treatment outcome were larger ($d \geqslant |1|$). When the true difference in treatment outcome was medium sized, the overall recovery performance was poor, especially for small sample sizes. One plausible reason for this phenomenon was sampling error: For example, for model A, when N equaled 200, the actual effect sizes (d) in the right leaves of the simulated data sets were far off -0.50 (range was -1.34 through 0.25). No difference in recovery performance was observed between situations with intercorrelation between the covariates ($\rho = 0.20$) and without intercorrelation. The total number of covariates influenced only the recovery of the structure of the true trees. When this number was higher ($J \geqslant 10$) , larger sample sizes were needed ($N \geqslant 500$) to obtain good recovery for more complex models. One plausible reason for this decline in performance is the problem of selection bias in tree-based methods, especially for data with many continuous covariates. A solution, as proposed by Loh [29], could be to consider in the selection process only the three quartiles (Q1, Q2, and Q3) as possible split points to compute the criterion C. Subsequently, the final split point can be computed optimally.

The recovery performance of QUINT was generally better than that of STIMA [17, Table 2, p. 521] and as good as Interaction Trees [18, Table 2, p. 149], for true models comparable in complexity and size of the interaction effect. It should be noted that the performance of QUINT was evaluated only for situations with true treatment–subgroup interactions based on a binary tree structure and not for situations with other types of true treatment–subgroup interactions, for example, cross products. This would be a challenge for future research.

4. Application to real data from Breast Cancer Recovery Project

4.1. Data

We re-analyzed data from the Breast Cancer Recovery Project for younger women with early-stage breast cancer [23, 30]. The majority of these women had had a lumpectomy and their axillary nodes removed

and had received combined radiation and chemotherapy. Within 2 months of having completed active nonhormonal adjuvant therapy, participants ($n = 252$) were randomly assigned to one of three therapy conditions: a control condition (standard medical care; $n = 84$), a nutrition intervention (how to adopt a low-fat, high-fruit, high-vegetable diet; $n = 85$), and an education intervention (information about breast cancer and provision of coping skills; $n = 83$). Our choice for these data was motivated by the fact that Scheier *et al.* reported results of well-performed moderated regression analyses [23]. This enabled us to compare QUINT with state-of-the-art moderator analysis. Previous analyses of these data showed that both the nutrition and education interventions were more effective than standard medical care [30] and that part of the main effect of intervention was moderated by patient baseline characteristics [23].

For our QUINT re-analyses of these data, we focused on the two nonstandard intervention groups and, therefore, excluded the patients in the control condition. We further used one outcome variable, for the construction of which we relied on a measure of depressive symptoms (by means of an abbreviated form of the Center for Epidemiologic Studies Depression Scale, see [23]). This measure was collected both at baseline and at a 9-month follow-up. We calculated a change score in such a way that a higher score referred to a better outcome. The resulting outcome variable was referred to as Improvement in depression (Table IV). Because of missing values at the 9-month follow-up, the available cases for analyses were 78 and 70, in the nutrition and education condition, respectively. As potential moderator variables, we included all numerical patient characteristics from the original study (measured at baseline) without missing values. Descriptive statistics for all variables that were included in our re-analyses are given in Table IV.

4.2. *Analysis with QUINT and results*

Given that the outcome variable was based on a Likert-type scale, QUINT analyses were performed using the treatment effect size criterion. Because the sample size was small ($N = 148$), a relatively large value of d_{min} ($d_{\text{min}} = 0.40$) was used in the qualitative interaction condition (Section 2.4.2) to check whether qualitative treatment–subgroup interactions were present in the data. A relatively large number of bootstrap samples ($B = 200$) was used to estimate the bias-corrected criterion values. The default parameter values were used for $L_{\text{upperlimit}}$ (i.e., $L_{\text{upperlimit}} = 10$), the weights of the partitioning criterion, and for the minimal sample size per treatment condition. After growing the full tree, we used the 1-SE

Table IV. Descriptive statistics for all variables involved in re-analyses of data from the Breast Cancer Recovery Project. The potential moderators were all measured at baseline (i.e., before receiving nutrition or education treatment).

			Nutrition $n = 78$	Education $n = 70$
Variable	Range		Mean (SD)	Mean (SD)
Potential moderators				
Age	30.5	51.4	44.4 (4.8)	43.8 (5.1)
Nationality (Caucasian vs. not)	0	1	0.91 (0.29)	0.96 (0.20)
Marital status (married vs. not)	0	1	0.67 (0.47)	0.83 (0.38)
Weight change (yes vs. no) [a]	0	1	0.44 (0.50)	0.46 (0.50)
Treatment extensiveness index[b]	−1.8	2.6	−0.1 (1.1)	0.1 (1.2)
Comorbidities[c]	0	13	3.0 (2.5)	2.4 (2.7)
Dispositional optimism[d]	6	24	16.4 (3.8)	17.5 (2.9)
Unmitigated communion[d]	15	42	29.7 (5.1)	29.1 (5.6)
Negative social interaction[d]	5	16	7.9 (2.4)	7.5 (2.2)
Outcome				
Improvement in depression	−12	18	2.4 (5.6)	0.7 (5.0)

SD, standard deviation.
[a]Weight gained or lost since diagnosis.
[b]A treatment extensiveness index was created by standardizing and aggregating type of surgery (1, lumpectomy; 2, mastectomy) with type of adjuvant treatment received (0, none; 1, radiation or chemotherapy; 2, both).
[c]Comorbidities was the sum of the checked potential comorbidities (e.g., diabetes, migraines, arthritis, or angina) and the reported conditions that the participant currently had (open question).
[d]These scales were measured by validated questionnaires (see for descriptions [23]).

pruning rule to assess the optimal tree size (as explained in Supplementary materials B), and we applied the validation procedure for small sample sizes (Supplementary materials C.2).

The QUINT analysis resulted in a full tree of seven leaves. The pruning rule indicated that the optimal size of the tree was five leaves (Figure 3). Before interpreting the pruned tree of QUINT, we inspected whether the bias-corrected estimates of C for each value of L (with $L = 2, \ldots, L_{\max}$) exhibited a clear maximum value. For this purpose, we plotted the bias-corrected estimates of C against L (Supplementary materials Figure 1). The plot showed a clear maximum at $L = 5$, suggesting that the bootstrap-based selection worked well here. The splits of the pruned tree (Figure 3) involved the variables Dispositional optimism, Amount of negative social interactions, and Treatment extensiveness. In the nodes of Figure 3, the effect sizes are expressed as the standardized difference of the mean of the nutrition intervention *minus* the mean of the education intervention. As a consequence, for the leaf assigned to $\wp_1$, the effect size d is positive, while for the leaves assigned to $\wp_2$, the effect size d is negative. Three subgroups of women were distinguished. One type of women (R_2, with effect size $d = 2.06$, and assigned to $\wp_1$) appeared to benefit more from a nutrition than from an education intervention; these were women whose condition was most severe from a psychological perspective (i.e., they were more pessimistic and reported receiving higher levels of negative social interaction) but who so far had received the least extensive physical treatment (i.e., lumpectomy without or with only one form of adjuvant therapy). Two types of women (assigned to $\wp_2$) constituted a subgroup that appeared to benefit more from education than from nutrition: one type (R_1, with effect size $d = -0.71$) consisting of more pessimistic women who were however OK in terms of social interaction and one type (R_4, with effect size $d = -0.96$) consisting of more (but not extremely) optimistic women. A third subgroup, for which both treatments were about equally effective, consisted of two types of women: the type (R_3) with the most severe condition, both from a psychological perspective and from the perspective of past treatment extensiveness, and a type (R_5) consisting of overly optimistic women.

In the validation procedure, we set the number of bootstraps (B) equal to 1000 and the value of $L_{\text{upperlimit}}$ equal to 5. The results showed that the mean range of the effect sizes in the leaves of the bootstrap solutions was 3.28. This mean range decreased to 1.77 in the leaves of the test solutions (using the original data as 'test data'). The resulting estimated bias (also called 'optimism') in the range of the effect sizes was 1.51. This result implies that caution is needed when generalizing the effect sizes in the leaves of Figure 3 to the entire population of younger women with early-stage breast cancer (also see Section 5).

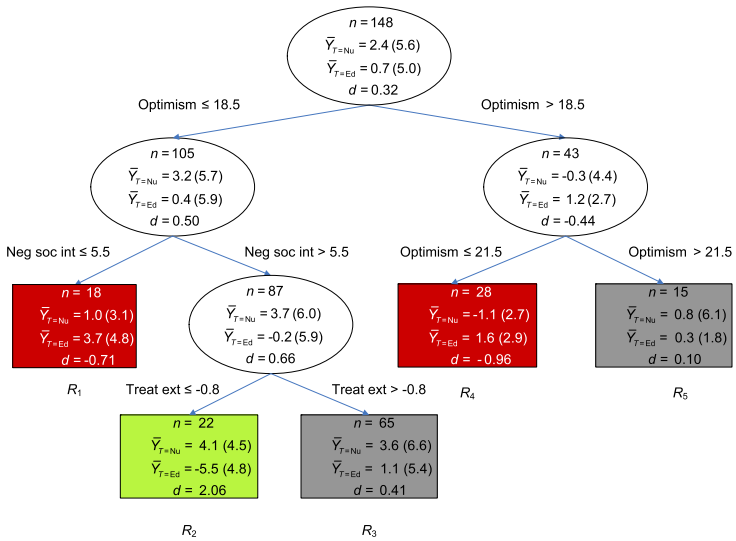

Figure 3. Results of application of QUINT to data from the Breast Cancer Recovery Project [23]. Pruned tree for Improvement in depression is shown. Each node contains the sample size (n), the conditional outcome means (and standard deviations) for the Nutrition and Education treatment groups ($\bar{Y}_{T=\text{Nu}}$ and $\bar{Y}_{T=\text{Ed}}$), and Cohen's effect size (d), expressed as the standardized mean difference between $T = \text{Nu}$ and $T = \text{Ed}$. Assignment of the leaves to the partition classes is represented in green, $\wp_1$; red, $\wp_2$; and dark grey, $\wp_3$. Optimism, Dispositional optimism; Neg soc int, Negative social interaction; Treat ext, Treatment extensiveness index.

4.3. Analysis with STIMA and Interaction Trees, and results

STIMA was applied to the data from the Breast Cancer Recovery Project using the R package 'stima' and Interaction Trees using R functions supplied to the authors by Su [18]. For both methods, the minimal node size was set at 10% of the sample size, and the maximum size of the tree was set at six leaves. For STIMA, we used leave-one-out cross-validation and the 0.60-SE rule to prune the tree. For Interaction Trees, we made use of 200 bootstrap samples that each played the role of test set, used $\lambda = 4$ for selecting the best-sized tree, and took the median best-size across all bootstrap samples.

The results of STIMA read that for both outcome variables, no treatment–subgroup interaction effects were present, with the regression trunks being pruned back to the root node. The solution of Interaction Trees was a tree with three leaves (Figure 4): two types of women for whom nutrition was considerably better than education (R_1, $d = 0.95$; R_3, $d = 1.50$) and one type for whom kind of treatment did not make a difference (R_2, $d = -0.14$). Obviously, this implies a quantitative treatment–subgroup interaction. Note that the first split of the resulting tree (Figure 4) is not allowed in QUINT, because of the qualitative interaction condition (Section 2.4.2).

4.4. Discussion

A comparison of our results with the results of the more classical approach to moderator analysis as applied in the original study of Scheier *et al.* [23] revealed that all moderator variables that were retained in the solutions by QUINT and Interaction Trees were also found by Scheier *et al.*, except for the Treatment extensiveness index. A plausible reason for this difference is that Scheier *et al.* focused only on two-way interactions, while according to the tree-based solutions, Treatment extensiveness appeared to contribute to higher-order interaction effects (Figures 3 and 4). Beyond the moderator variables that were selected, QUINT and Interaction Trees provided several distinctive advantages over the classical approach to moderator analysis. First, they resulted in a large reduction of the number of required analyses. Second, they yielded additional information with regard to the cutoff points on the moderators that are important for treatment–subgroup interactions. Last but not least, they provided an insightful (and substantively meaningful) picture of the overall pattern of moderation for the whole of all relevant moderator variables. Technically speaking, this also involved the detection of higher-order interactions and nonmonotonic relationships between treatment and baseline characteristics of the patients (e.g., a nonmonotonic relationship was found between difference in treatment outcome and Optimism, see Figure 3). Such interactions are typically hard to retrieve in standard moderator analyses.

In contrast to the results of QUINT and Interaction Trees, STIMA did not detect any treatment subgroup–interaction effects. The most plausible reason for this is that STIMA also includes main effects in the model, which may hamper the detection of interaction effects, especially in smaller data sets. Even

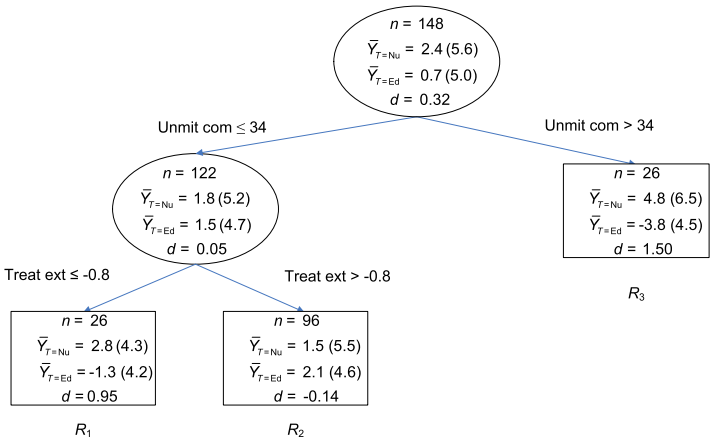

Figure 4. Result of application of Interaction Trees to data from the Breast Cancer Recovery Project [23]. Pruned tree for Improvement in depression is shown. Each node contains the sample size (n), the conditional outcome means (and standard deviations) for the Nutrition and Education treatment groups ($\bar{Y}_{T=\text{Nu}}$and $\bar{Y}_{T=\text{Ed}}$), and Cohen's effect size (d), expressed as the standardized mean difference between $T = Nu$ and $T = Ed$. Unmit com, Unmitigated communion; Treat ext, Treatment extensiveness index.

more importantly, the solutions of QUINT and Interaction trees were different (Figures 3 and 4), with respect to the size of the tree, the first splitting variable, and the partitioning of the observations into subgroups (i.e., Cohen's κ was 0.28). The solution of Interaction Trees was dominated by a quantitative interaction effect, whereas QUINT revealed a qualitative treatment–subgroup interaction with clear implications for optimal treatment assignment: On the basis of the (small) marginal treatment effect (i.e., the intervention main effect), a plausible decision would be to assign all future patients to the nutrition intervention. The expected treatment benefit (i.e., decrease in depression) then would be 2.4 (instead of 0.7 with the education intervention; see root nodes of Figures 3 and 4). On the basis of the QUINT solution, one would assign the patients with characteristics that lead to end node R_2 (15% of the population; Figure 3) to the nutrition intervention; their expected treatment benefit would be 4.1. Furthermore, one would assign the patients with characteristics that lead to the end nodes R_1 and R_4 (31% of the population; Figure 3) to the education intervention; their expected treatment benefit would be, respectively, 3.7 and 1.6. For the patients in R_3 and R_5, there is no clear benefit for one treatment over the other, and assignment could be carried out on other grounds, for example, accessibility of the treatment. (Note that some caution is needed here: To determine the expected treatment benefit for future patients, the ideal situation would be to estimate the treatment outcome in the leaf nodes of the trees for an independent test data set, see Supplementary materials C.)

5. General discussion

In the present paper, we dealt with the ubiquitous finding of treatment effect heterogeneity. In many contexts, there is an increasing awareness that a 'one size fits all' approach to treatment may be far from optimal. Hence, the construction of suitable methodological tools for the identification and study of treatment–subgroup interactions is of high pragmatic importance, not in the least in view of the development of evidence-based personalized medicine.

In the construction of a suitable methodology for the study of treatment–subgroup interactions, one may not ignore a research setting that shows up very often in empirical practice, namely that of RCTs without strong a priori hypotheses on the nature of subgroups involved in treatment–subgroup interactions. Two kinds of settings may arise: a setting with a small number of potential treatment effect modifiers (for which advanced methods such as fractional polynomials are suitable [12]) and a setting with a large number of potential effect modifiers. At present, typical methods that are available for the latter setting are of a tree-based sequential partitioning type. In this paper, we reviewed a few of these methods, and we proposed a novel member of the family, QUINT.

An advantage of the tree-based methods under study is that they yield output that lends itself (in principle) to a straightforward interpretation [27].[¶] The relevance of the methods for treatment–subgroup interactions is further obvious when looking at their associated objective functions in which such interactions play a key role. The distinctive contribution of QUINT in this regard is its focus on *qualitative* treatment–subgroup interactions. These are especially important if the primary concern of the researcher is on optimal treatment assignment. In fact, the objective function of QUINT is a straightforward formalization of a qualitative interaction with significant pragmatic consequences on a treatment assignment level.

Tree-based methods (including QUINT) can efficiently deal with RCT data that include large numbers of baseline characteristics. A major reason for this is their stepwise, greedy nature. Importantly, this efficiency may be at the expense of ending up in locally rather than globally optimal solutions (although, the simulation results of QUINT point at a rather satisfactory optimization performance). For instance, in each step of QUINT, a split involving an empty $\wp_1$ or $\wp_2$ is discarded. However, by discarding such splits, it is possible that in the next step, a more powerful split in terms of the partitioning criterion will be missed.

Beyond any doubt, the Achilles heel of post hoc methods to derive subgroups from RCT data that are involved in treatment–subgroup interactions is the risk of inferential errors. This includes failures to detect interactions and especially also identifying apparent interactions that cannot be replicated in

[¶]*Otherwise, the easy interpretability of trees immediately relates to the fact that most tree-based methods (including QUINT) focus on rectangular partitions (i.e., partitions of the form $X_1 > s_1 \wedge X_2 > s_2$, where s_1 and s_2 denote split points). If one conjectures that weighted linear combinations of baseline characteristics are important moderators and if one wants to induce such combinations during the data analysis, the use of other methods is advisable (e.g.,* [31]*). If such combinations are known a priori, they can obviously be included in a tree-based analysis.*

follow-up studies [1, 30]. Whereas the latter is a bottleneck for all post hoc methods, tree-based methods are especially vulnerable at this point as they rely on a very large search space based on a very huge number of covariate split-point combinations. To be sure, several tree-based methods (including QUINT) make use of cross-validation-based pruning procedures to prevent themselves from overfitting the data at hand, but this cannot be considered a foolproof solution. With regard to QUINT, in the present paper, an attempt has been made to deal with the problem of inferential errors in a constructive way by subjecting such errors to a systematic investigation in a set of simulations. It seems fair to conclude that these simulation yielded mixed results: On the one hand, they gave a good indication of types I and II error rates under several conditions (these rates being comparable with that of other tree-based methods), and they also provided clues about controlling these levels through the choice of a suitable critical value for an initial test of absence of qualitative interaction. On the other hand, finding a good balance between types I and II error rates appeared to be rather tricky. Moreover and most importantly, to arrive at acceptable rates large sample sizes (and probably larger than feasible within several contexts) seem to be inevitably necessary. Probably the safest way to deal with QUINT and related tree-based methods is to consider them exploratory tools that yield useful hypotheses. Subsequently, these hypotheses should be tested in follow-up confirmatory research with new RCTs that make use of a stratified randomization in which the strata are constructed on the basis of the splitting variables and split points as identified by QUINT.

Acknowledgements

The research reported in this paper was supported by the Flemish Fund for Scientific Research (G.0546.09), the Belgian Federal Science Policy (IAP P7/06), the Research Fund of K.U. Leuven (GOA/2010/02), and the Netherlands Organization for Applied Scientific Research TNO. For the simulations, we used the infrastructure of the VSC – Flemish Supercomputer Center, funded by the Hercules foundation and the Flemish Government – department EWI. The authors gratefully acknowledge Michael Scheier for making available the data for the empirical application, Xiaogang Su for supplying his R code, and Eva Ceulemans, Lisa Doove, and Stef van Buuren for their helpful comments on previous versions of this paper.

References

1. Rothwell PM. Treating individuals 2. Subgroup analysis in randomised controlled trials: importance, indications, and interpretation. *Lancet* 2005; **365**:176–186.
2. Byar DP. Assessing apparent treatment-covariate interactions in randomized clinical trials. *Statistics in Medicine* 1985; **4**:255–263.
3. Kraemer HC, Wilson GT, Fairburn CG, Agras WS. Mediators and moderators of treatment effects in randomized clinical trials. *Archives of General Psychiatry* 2002; **59**:877–883.
4. Lubin A. The interpretation of significant interaction. *Educational and Psychological Measurement* 1961; **21**:807–817. DOI: 10.1177/001316446102100406.
5. Mokken RJ, Lewis C. A nonparametric approach to the analysis of dichotomous item responses. *Applied Psychological Measurement* 1982; **6**:417–430.
6. Behrendt CE, Gehan EA. Treatment–subgroup interaction: an example from a published, phase II clinical trial. *Contemporary Clinical Trials* 2009; **30**:279–281. DOI: 10.1016/j.cct.2009.02.002.
7. Tunis SR, Benner J, McClellan M. Comparative effectiveness research: policy context, methods development and research infrastructure. *Statistics in Medicine* 2010; **29**:1963–1976.
8. Shuster J, Van Eys J. Interaction between prognostic factors and treatment. *Controlled Clinical Trials* 1983; **4**:209–214.
9. Schaffer JP. Probability of directional errors with disordinal (qualitative) interaction. *Psychometrika* 1991; **56**:29–38.
10. Cohen J, Cohen P, West SG, Aiken LS. *Applied Multiple Regression/Correlation Analysis for the Behavioral Sciences*, 3^{rd} edn. Lawrence Erlbaum: Mahwah, NJ, 2003.
11. Cronbach L, Snow R. *Aptitudes and Instructional Methods: A Handbook for Research on Interactions*. Irvington Publishers: New York, 1997.
12. Royston P, Sauerbrei W. A new approach to modelling interactions between treatment and continuous covariates in clinical trials by using fractional polynomials. *Statistics in Medicine* 2004; **23**:2509–2525. DOI: 10.1002/sim.1815.
13. Gelman A, Huang Z. Estimating incumbency advantage and its variation, as an example of a before-after study. *Journal of the American Statistical Association* 2008; **103**:437–451. DOI: 10.1198/016214507000000626.
14. Feller A, Holmes CC. Beyond toplines: heterogeneous treatment effects in randomized experiments. *Technical Report*, University of Oxford, Oxford, UK, 2009. Available from: http://www.stat.columbia.edu/~gelman/stuff_for_blog/feller.pdf [Accessed on August 30, 2011].
15. Wang R, Lagakos SW, Ware JH, Hunter DJ, Drazen JM. Statistics in medicine: reporting of subgroup analyses in clinical trials. *New England Journal of Medicine* 2007; **357**:2189–2194.
16. Dusseldorp E, Meulman JJ. The regression trunk approach to discover treatment covariate interaction. *Psychometrika* 2004; **69**:355–374.

17. Dusseldorp E, Conversano C, Van Os BJ. Combining an additive and tree-based regression model simultaneously: STIMA. *Journal of Computational and Graphical and Statistics* 2010; **19**(3):514–530. DOI: 10.1198/jcgs.2010.06089.
18. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research* 2009; **10**:141–158.
19. Su XG, Zhou T, Yan X, Fan J, Yang S. Interaction trees with censored survival data. *The International Journal of Biostatistics* 2008; **4**(1). Article 2. Available from: http://www.bepress.com/ijb/vol4/iss1/2 [Accessed on July 22, 2011].
20. Foster JC, Taylor JMG, Ruberg SJ. Subgroup identification from randomized clinical trial data. *Statistics in Medicine* 2011; **30**:2867–2880. DOI: 10.1002/sim.4322.
21. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect research – a recursive partitioning method for establishing response to treatment in patient populations. *Statistics in Medicine* 2011; **30**:2601–2621.
22. Doove L, Dusseldorp E, Van Deun K, Van Mechelen I. A comparison of five sequential partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. Manuscript submitted for publication.
23. Scheier MF, Helgeson VS, Schulz R, Colvin S, Berga SL, Knapp J, Gerszten K. Moderators of interventions designed to enhance physical and psychological functioning among younger women with early-stage breast cancer. *Journal of Clinical Oncology* 2007; **25**:5710–5714. DOI: 10.1200/JCO.2007.11.7093.
24. Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2^{nd} edn. Lawrence Erlbaum: Hillsdale NJ, 1988.
25. Efron B. Estimating the error rate of a prediction rule: improvements on cross-validation. *Journal of the American Statistical Association* 1983; **78**:316–331.
26. LeBlanc M, Crowley J. Survival trees by goodness of split. *Journal of the American Statistical Association* 1993; **88**:457–467.
27. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Chapman & Hall/CRC: Boca Raton, 1984.
28. R Developmental Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing: Vienna, 2012.
29. Loh WY. Regression trees with unbiased variable selection and interaction detection. *Statistica Sinica* 2002; **12**:361–386.
30. Scheier MF, Helgeson VS, Schulz R, Colvin S, Berga S, Bridges MW, Knapp J, Gerszten K, Pappert WS. Interventions to enhance physical and psychological functioning among younger women who are ending nonhormonal adjuvant treatment for early stage breast cancer. *Journal of Clinical Oncology* 2005; **23**:4298–4311. DOI: 10.1200/JCO.2005.05.362.
31. Kraemer HC. Discovering, comparing and combining moderators of treatment on outcome after randomized controlled trials: a parametric approach. *Statistics in Medicine* 2013; **32**:1964–1973. DOI: 10.1002/sim.5734.