Subgroup identification in clinical trials: Comparing methods QUINT and OTR with a focus on subgroups with no treatment difference

Chen, T.J.

Subgroup identification in clinical trials

Comparing methods QUINT and OTR with a focus on subgroups with no treatment difference

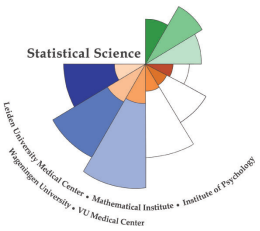
Ting-Jung, Chen (s2047500)

Thesis supervisor:
Dr. Elise Dusseldorp
Dr. Marjolein Fokkema

MASTER THESIS

Defended on November, 2019

Specialization: Data Science



STATISTICAL SCIENCE
FOR THE LIFE AND BEHAVIOURAL SCIENCES

Abstract

In clinical trials, heterogeneity of treatment effect often exists between patients with different pretreatment characteristics, such as age, gender, weight, etc. In response to such issue, various subgroup identification approaches have been proposed. Two methods among them, Qualitative Interaction Tree (QUINT) and a method adapted from an optimal treatment regimes (OTR) approach proposed by Zhang et al. (2012), are compared in this paper. These two methods identify three types of subgroups in a situation with two treatments (A and B): one subgroup for which treatment A is better than treatment B, one for which treatment B is better than treatment A, and one for which the difference between the two treatment outcomes is negligible (called "indifference group").

A simulation study was conducted to compare the two methods with regard to their recovery performance (quantified by type I error rates, type II error rates, Cohen's κ agreement to the true subgroups, and splitting performance of the derived trees) and their predictive performance (quantified using the difference between the true expected treatment outcome and the estimated treatment outcome of sample data and population data). Results of the simulation study suggested that QUINT has its advantage in recovering the subgroups, and the method adapted from the OTR approach has its advantage in predicting treatment outcome.

Keywords— subgroup analysis, qualitative treatment-subgroup interaction, indifference group, decision tree

1 Introduction

The presence of treatment-effect heterogeneity has always brought up critical consideration in clinical studies. With the recent attention gained in personalized medicine, it has become a popular topic to investigate how treatments affect differently on different patients. Who gains a large treatment effect and who gains a small effect? For whom is the treatment beneficial and for whom harmful? Formally, a "treatment effect" is defined as a measure comparing one treatment to another treatment (or no treatment). For instance, in randomized controlled trials (RCT), it indicates the difference in outcome between a treatment condition and a no treatment condition or a standard care condition, etc. To explore differential treatment effectiveness, methods to identify subgroups that differ in treatment effect are no doubt a critical core of interest. With the help of proper identification, subgroups can be treated differently based on their own characteristics and consequently gain more efficacy compared to a one-for-all approach.

Recently, Lipkovich et al. published a nice overview study about data-driven treatment subgroup identification methods [4]. Among the approaches, some focus on identifying subgroups of patients for whom treatment A is better than B, and vice versa (subgroups for whom B is better than A), referred as qualitative treatment-subgroup interactions, suggesting that the treatment effect differs between subgroups not only in magnitude but also in the direction of the effect; and some approaches emphasize on recovering optimal treatment regimes (OTR), that are the treatment assignment rules which give subgroups of patients their optimal treatment. Qualitative Interaction Trees (QUINT; Dusseldorp et al., 2014) and an OTR method proposed by Zhang et al. (2012) are two representative methods of these two kinds of approaches accordingly.

In this paper, QUINT is compared with Zhang's OTR approach (2012) in situations with two treatments (A and B). QUINT adapts the regular binary partitioning algorithm of classification and regression trees (CART, Breiman et al., 1984) in such a way that the subgroups of interest are identified, whereas Zhang's OTR approach first computes a contrast value for every patient, which quantifies the relative expected benefit of receiving treatment B over treatment A, and then adopts these contrast values on CART using its regular binary partitioning algorithm. Because both of them estimate tree-based treatment regimes, the implementation on CART made Zhang's OTR approach intriguing to compare with QUINT. The main difference between them is that QUINT also discovers subgroups with negligible difference in treatment outcome between A and B (the so-called indifference groups) when growing the trees. The method proposed by Zhang et al. does not take indifference groups into account. However, recognizing indifference groups is an important issue with practical relevance. For instance, theoretically speaking, patients would choose treatment A while being told that treatment A works better than treatment B on him/her. Yet, when the effect of treatment A is not remarkable enough, the patient may still choose treatment B over treatment A due to its better accessibility, such as a lower price, a closer clinic providing treatment B, etc.

To investigate how to recognize a "remarkable" effect, an interaction probing technique was proposed by Johnson and Neyman (J-N technique; Johnson & Neyman, 1936). This technique aims at defining *a region of insignificance* and states that when an interaction effect appears between a focal predictor X (e.g., a treatment variable) and a moderator variable z (e.g., a patient characteristic) on an outcome variable y, it should only be recognized as having significant meaning when the value of the moderator does not fall in the *region of insignificance*. To illustrate in Figure 1(a), only under certain value of the moderator z (i.e., $< z_1$ or $> z_2$), will the outcome y differ between the categories of X. This technique is pre-eminently suited for estimating trees with three types of subgroups. Figure 1(b) outlines how the three subgroups are defined on the values of z in a tree. This further contributes to the study of qualitative treatment-subgroup interaction with indifference group being considered, where

the patient's characteristics play as the role of the moderator; the treatment is the focal predictor and the indifference group corresponds to the insignificance region in J-N technique.

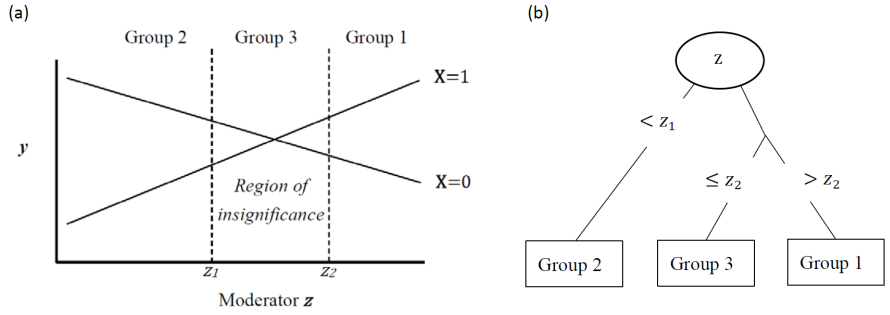


Figure 1: Graphic (a) in the left panel shows that there is an interaction effect between the focal predictor X and the moderator z on the outcome y. An insignificance region between z_1 and z_2 is defined by J-N technique. When the observed z falls in this region, the interaction between focal predictor X and outcome y is concluded as without significant meaning. This region serves as a similar role of indifference group in the method of qualitative treatment-subgroup identification, indicating that the effect on y under neither category of X is superior. In graphic (b) shows the corresponding tree of (a), which defined the three subgroups on z.

In a study held by Sies and Van Mechelen (2017), QUINT was compared with the OTR approach by Zhang et al. but with the assignment of the indifference group being disregarded by re-assigning such leaves to the treatment with the highest mean outcome in that leaves [6]. Nonetheless, to compare these two approaches, one may also determine the indifference group post-hoc for Zhang's OTR approach based on its terminal leaves instead of dismissing the indifference group assignment in QUINT. The details of the post-hoc assignment will be presented in a later section. Here in this paper, we will focus on QUINT and Zhang's OTR with post-hoc assignment (PostZhang) and their performance on recovering the subgroups as well as their predictive performance in terms of the estimated treatment outcome for sample data and the population. Because the algorithm of QUINT takes indifference groups into account while growing the tree, we hypothesize that the performance of recovering the true subgroups by QUINT is better than that by Zhang's OTR in true scenarios including indifference groups.

The structure of the remainder of this paper is as follows: In Section 2, QUINT and Zhang's OTR approach (2012) with post-hoc indifference group assignment will be introduced. The simulation study for the comparison of these two methods will be outlined in section 3 and the results are reported in section 4. The corresponding discussion will be carried out in Section 5.

2 Methods

2.1 Qualitative interaction trees (QUINT)

The method of QUalitative INteraction Trees (QUINT) was first introduced by Dusseldorp et al. (2014). It is a method targeted on detecting qualitative treatment-subgroup interaction as well as which variables contribute to such interaction through a binary tree. The tree aims to partition the study population into three groups signified as P_1, P_2 and P_3. P_1 represents a subgroup of patients gaining larger treatment outcome via treatment A than treatment B and P_2 is a subgroup having the situation the other way around (treatment B better than treatment A). As for P_3, known as the indifference group, the patients react with no notable difference between either treatment.

The binary tree starts with a root node containing all the patients and performs binary splits recursively on some patients' characteristics that can maximize the partitioning criterion (C) until no larger criterion value can be found or reaching one of the other stopping criteria, that is when P_1 and P_2 do not exist simultaneously after the first split; or the number of subjects assigned to treatment A or B is smaller than the predefined number.

The partitioning criterion, C, consists of two components: *Difference in treatment outcome* and *Cardinality*. These two components ensure that each time the split is chosen in a way such that the difference between outcomes via treatment A and B in created subgroups P_1 and P_2 are as large as possible, along with comparable sample sizes in both groups. Criterion C is defined as

$$C = w_1[log(1 + D_1) + log(1 + D_2)] + w_2[log(N_1) + log(N_2)], \tag{1}$$

where D_1 and D_2 express the weighted average of difference in treatment outcome across all the current terminal nodes belonging to P_1 and P_2 and together compose the *Difference in treatment outcome* component. N_1 and N_2 represent the number of subjects belong to partition class P_1 and P_2 and consist the *Cardinality* component. These two components are weighted by pre-defined weights, w_1 and w_2. By optimizing this partitioning criterion, the qualitative treatment-subgroup interaction with largest possible practical significance can be identified. [3]

2.2 Optimal Treatment Regimes

2.2.1 Original Method by Zhang et al.(2012)

The estimation of optimal treatment regimes (OTR) aims at exploring the assignment rule that assigns a treatment that works best among a set of possible treatments,

treatment A and treatment B in our case, to a patient based on his/hers own characteristics [7]. In 2012, Zhang et al. introduced a method that transformed the problem of estimating OTR with a two treatments setting into a weighted classification problem. They achieved this by adapting the direction and the magnitude of the estimated contrast value to the label of class and the cost of being misclassified respectively, on top of the general framework of OTR approaches where the contrast value of potential outcomes under the two treatments is estimated. It is noted that the potential outcomes are adopted from the potential outcome framework of Rubin (1978) and denote the possible outcomes a patient would have under different treatment conditions [5].

The contrast value estimator applied in Zhang's method is referred as a doubly robust augmented inverse probability weighted estimator (AIPWE), which is defined as follow:

$$\hat{C}_{AIPWE}(X_i) = \frac{T_i}{\pi(X_i,\hat{\gamma})}Y_i - \frac{1-T_i}{1-\pi(X_i,\hat{\gamma})}Y_i - \frac{T_i-\pi(X_i,\hat{\gamma})}{\pi(X_i,\hat{\gamma})}\mu(T_i=1,X_i,\hat{\beta}) - \frac{T_i-\pi(X_i,\hat{\gamma})}{1-\pi(X_i,\hat{\gamma})}\mu(T_i=0,X_i,\hat{\beta}), \quad (2)$$

where the subscript i represents data at patient's level, X denotes the pre-treatment characteristics, T denotes the treatment being assigned with value 0 signifying the assignment to treatment A and 1 to treatment B, and Y denotes the obtained treatment outcome; $\pi(X;\hat{\gamma})$ denotes the estimated propensity score. In our study it is estimated by the sample proportion of getting treatment B (i.e., $P(T=1)$), because we consider only randomized trials. $\mu(T_i, X_i, \hat{\beta})$ denotes a model for potential outcome under treatment T_i, here a linear regression model on Y using T, X and their interaction term as the independent variables is implemented. Note that the property of AIPWE makes the estimators robustness against the misspecification of this regression model.

With these estimated contrast values at hand, the OTR problem is transferred to a classification problem via the mechanism mentioned in the beginning of this section. That is, the direction of a contrast values is viewed as a class label, defined as $\hat{Z}_i = I\{\hat{C}_{AIPWE}(X_i) > 0\}$; and the magnitude, $|\hat{C}_{AIPWE}(X_i)|$, is regarded as the cost of the subject being misclassified. The weighted classification problem with the class labels as outcomes, the costs of each subject being misclassified as case weights, and the characteristics as splitting candidates (predictors) is then solved utilizing CART. Consequently, each patient is classified to one of the two classes with label 1 and 0. Concerning our scheme with 2 treatments, A and B, we identify the class with label 0 as a subgroup of patients benefits more from treatment A than B and the class with label 1 as a subgroup that benefits more from treatment B than A. It is worth mentioning that these two classes coincide with the partition classes P_1 and P_2 in QUINT and thus are also referred to as P_1 and P_2 in the remaining of the paper. In

the next subsection, we further propose an approach that additionally yields a class that coincides with the QUINT partition class P_3 from Zhang's OTR method.

2.2.2 Zhang's Method with Post-Hoc Assignment (PostZhang)

With the use of the method proposed by Zhang et al. (2012), we are able to estimate regimes that assign patients to one of the two treatments that is believed to benefit the patients more, depending on their own profile. Yet, it is also valuable to identify when the benefit is ambiguous and makes neither one be practically superior, just as the indifference group P_3 in QUINT. Hence, we extend the work of Zhang and accommodate an approach to reassign the terminal leaves (nodes) of trees obtained by the Zhang's method with the indifference group being considered. Since the assignment to the indifference group is done post-hoc, we further refer to this extended method as "PostZhang" in this paper.

On top of the original method proposed by Zhang, which uses the AIPWE estimators for contrast values and CART for the classification, we introduce *Cohen's d effect size* (standardized mean difference) as a measure to further justify whether a terminal leaf (node) yielded from Zhang's method should be reassigned to the indifference group. The standardized mean difference for leaf l, d_l, is defined as follows:

$$d_l = \frac{(\bar{Y}_{T=0} - \bar{Y}_{T=1})}{s_l} \tag{3}$$

$$s_l = \frac{\sqrt{(n_0 - 1)s^2_{T=0,l} + (n_1 - 1)s^2_{T=1,l}}}{n_0 + n_1 - 2}, \tag{4}$$

where n_0 and n_1 respectively denote the sample size for $T = 0$ and $T = 1$ in leaf l. As above defined, d_l quantifies the difference in treatment outcomes by mean of its pooled estimate of the population standard deviation of the treatment groups in the leaf. A leaf is concluded as having ambiguous benefit from either treatment and should be reassigned to the indifference group if its d_l is smaller than a predefined threshold.

The threshold is a parameter that can be defined by users based on their own research field. The choice of it usually subjects to a variety of factors, both theoretically and practically, such as the disorder itself, study objectives, and treatment accessibility, etc. However, there is a general guideline proposed by Cohen in 1988, which is commonly recommended when interpreting the standardized mean difference (effect size). Cohen stated that an effect size of 0.2, 0.5 and 0.8 should be accordingly considered as "small effect", "medium effect" and "large effect" [2]. To be more specific, a "small effect", i.e. $d = 0.2$, should be a good choice as threshold such that the subgroups of patients having small treatment effect can be regarded as indifference groups. Yet, the choice of the threshold may vary in our simulation study, and will be outlined in the following section.

3 Simulation study

We conducted a simulation study to compare the performance of QUINT and PostZhang. Performance was assessed both in terms of recovery of true treatment subgroups (called Recovery Performance), and in terms of the predicted treatment outcome (called Predictive Performance). Before comparing these two methods, we first investigated each of them with different specification of tuning parameters and elaborated the choice of those parameters based on analyses which explored the performance of recovering subgroups.

3.1 Design

This simulation study was conducted on artificial data sets. These data sets resemble the scheme of a randomized trial, under which N patients are treated by either treatment A or treatment B. There are J pretreatment characteristics $(X_1, X_2, \ldots, X_J)$ for every patient. These characteristics are continuous variables that distribute multivariate-normally with $\mu_{X_j} = 0, \sigma_{X_j} = 1$ for all variables and the covariance $\sigma_{j,j'} = \rho$ for any two variables. The design factors with the levels that were manipulated were:

- *Sample size* N: 300 and 1000
- *Number of characteristics* J: 5 and 20
- *Correlation between characteristics* ρ: 0 and 0.4
- *Treatment effect in the leaves* d: 0.5 (small), 1 (medium) and 2 (large)

Given the design factors, the treatment outcome Y_i for patient i with characteristics $\mathbf{X}_i = (X_{i1}, X_{i2}, \ldots X_{iJ})$ is formulated as follows:

$$Y_i = 1 + 0.25X_{i1} + 0.25X_{i2} - 0.25X_{i5} - d[1 - g^{P3}(\mathbf{X}_i)][T_i - g^{opt}(\mathbf{X}_i)]^2 + \epsilon_i, \quad (5)$$

where a higher value in Y_i is defined as having a better treatment outcome. The error term ϵ is standard normally distributed. T_i is a binary variable denoting the treatment being assigned to. It follows a Bernoulli distribution with $\theta = 0.5$, and with $T_i = 0$ implying treatment A and $T_i = 1$ implying treatment B. Function $g^{opt}(X)$ signifies the optimal treatment given $\mathbf{X}$, while function $g^{P3}(X)$ denotes the indicator for being in indifference group. This data structure is modified from an initial model proposed by Sies [6]. We revised the model by adding a term with the P_3 indicator, g^{P3}, to accommodate the scenario with an indifference group being considered. As illustrated in (5), if a patient belongs to this indifference group, his/her treatment outcome will not be affected whether he/she is assigned to the optimal treatment or not. Contrarily, if a patient does not belong to the indifference group and is assigned to a non-optimal treatment according to his/her characteristics, the treatment outcome will be penalized by the design factor d.

The two models used in our simulation study are further determined by the specification of function g^{opt} and g^{P3} in (5) :

- Model 1:
  $g^{opt}(\mathbf{X}) = I(X_1 > -0.433) \times I(X_2 < -0.219)$
  $g^{P3}(\mathbf{X}) = I(X_1 > -0.433) \times I(X_2 \geq -0.219)$
- Model 2:
  $g^{opt}(\mathbf{X}) = 1$
  $g^{P3}(\mathbf{X}) = I(X_1 \leq -0.25)$,

where $I(.)$ is an indicator function. The split points for each splitting variable in Model 1 are designed in such a way that the three subgroups (P_1, P_2 and P_3) will have comparable sizes. Whereas in Model 2, the split point of X_1 in P_3 indicator is intended to get 2/5 of the sample being in P_3. It is noteworthy that, as shown in Figure 2, Model 2 reveals a scenario without a qualitative treatment-subgroup interaction in which Treatment B is always the better choice for patients who do not belong to the indifference group. This model was implemented to test the type I error rates of probing a qualitative interaction. (See subsection 3.3 for more details).

Using a fully crossed factorial design, we obtained $2 \times 2 \times 2 \times 3 = 24$ design cells. For each of them, we generated 100 data sets and brought out 2,400 data sets in total for each model.

3.2 Analysis

As abovementioned, the simulation study was performed in two stages. Firstly, the tuning parameters were investigated within each of the methods separately to decide the optimal setting. Secondly, the two methods were compared with parameters settings as determined in the first stage. In the first stage, we focused on evaluating *Recovery Performance* and in the second stage *Predictive Performance* was additionally inspected. See subsection 3.3 for details on these evaluation criteria.

In the first stage of the analysis, we examined QUINT with various values of the tuning parameter, *dmin*. It is a parameter used in QUINT method to check whether a qualitative interaction (i.e., the qualitative interaction condition) is present in a pruned tree. It is defined as the minimum value of standardized mean difference (Cohen's d effect size) that should hold true in at least one leaf of both P_1 and P_2. According to results from the simulation study by Duseldorp et al. (2014), *dmin*=0.3 can generally strike a good balance between type I error and type II error under sufficient sample sizes ($N > 300$). When a data with a small sample size is of interest, a higher *dmin* may be considered. Nonetheless, the recommendation was made on a previous implementation of *dmin*, where *dmin* was applied right after the first split to check for the qualitative interaction. The latest implementation has been changed to utilize

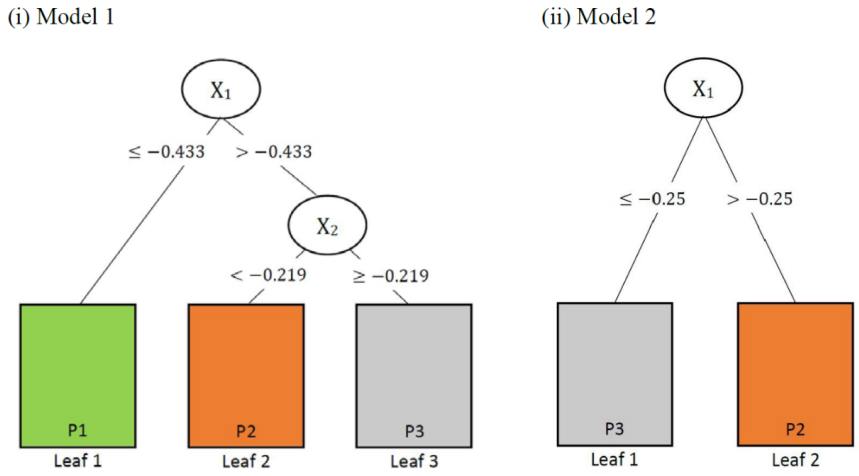


Figure 2: The true models where the simulated data set are generated from. Left: Model 1. Right: Model 2. The partition groups are labeled in the terminal leaves with P_1 colored in green and represents a subgroup of patients benefits more from Treatment A than Treatment B. Group P_2 is colored in orange and has the situation the other way around (Treatment B better than Treatment A). The leaves colored in gray denote the indifference group P_3, which shows no notable difference between two treatment outcomes. The labels of leaves are numbered ascendingly from left to right.

after trees are pruned. To investigate this new implementation, both *dmin*=0.3 and *dmin*=0.4 with respect to their Recovery Performance were inspected.

As for PostZhang, we investigated different values of the threshold parameter, *threshold.d*, which we constructed to determine whether a leaf should be assigned to the indifference subgroup (P_3) post-hoc based on its Cohen's d effect size. As the general guideline proposed by Cohen (1988) stated that an effect size of 0.2 is an effect with small size, we considered 0.2 as a preferred threshold to categorized those subgroups with small effect to subgroup P_3. However, there are ambiguous definitions for the effect size between 0.2 (small effect) and 0.5 (medium effect). To investigate it more thoroughly, the threshold with values 0.2, 0.3 and 0.4 [1] were inspected. In addition, the method of PostZhang with a zero-threshold was investigated which refers to the original method of Zhang. The method with these values of the threshold were inspected considering their Recovery Performance.

We applied repeated-measures analysis of variance (ANOVA) on the evaluations to

[1] Although the tuning parameters in QUINT and PostZhang are both measuring the *Cohen's d effect size* of leaves. The different definitions of the tuning parameters make the values being tested different between the two methods. Parameter *dmin* is the minimum value of the leaf effect size in at least one leaf of P_1 and P_2 separately in QUINT. The threshold parameter in PostZhang defines the minimum effect size for all the leaves classified to P_1 and P_2.

determine the specification among these values of the tuning parameter. In this stage, the ANOVA analyses were applied within each method (i.e., QUINT and PostZhang) using the evaluation outcomes of Recovery Performance as dependent variables, the specification of the tuning parameter as the within-subject variable, and the design factors as the between-subject variables. All the interaction terms between independent variables were included. The ANOVA results were interpreted in terms of generalized eta squared (η^2_G) instead of p-value due to the well-known problem of p-value in large samples. In the simulation study, data with size of 2,400 was applied for each ANOVA analysis. This large size may induce many trivial results with p-value lower than 0.05. Alternatively, a threshold of 0.02 on η^2_G was applied according to the guideline proposed by Bakeman (2005)[2]. Variables were concluded as having leading influence on the evaluation outcomes if they yielded η^2_G that was larger than 0.02.

Subsequently, the second stage of the analysis focused on the comparison between the two methods, QUINT and PostZhang, regarding their Recovery Performance and Predictive Performance. The two methods were applied with the setting of tuning parameters derived in the first stage. Also here, ANOVA analyses were performed on each evaluation outcome to illuminate the leading factors that influenced the outcomes. The ANOVAs were applied with *method* (i.e., QUINT or PostZhang) as the within-subjects variable and the four design factors (i.e., *sample size*, *number of characteristics*, *correlation between characteristics* and *treatment effect size*) as the between-subjects variables. All the interaction terms between these independent variables were included. Again, as in the first stage, the variables with $\eta^2_G > 0.02$ were concluded as the factors that influenced the evaluation outcomes.

In the method of QUINT, the data sets were analyzed using R-package *quint* (version 2.0). The *Difference in treatment outcome* component of the partitioning criterion was calculated in terms of Cohen's d effect size (`crit="es"`). The maximum number of leaves that can be reached when growing trees was set as `maxl=6` and `maxl=4` for data sets generated from Model 1 and Model 2, respectively.

In the method of PostZhang, the data sets were analyzed using our own code in R and R-package *rpart* with 40 as the minimum number of patients for a node to be split set (`minsplit=40`), 20 as the minimum number of patients in any terminal leaves (`minbucket=20`) and four as the maximum depth of any terminal leaf (`maxdepth=4`). The default complexity parameter of 0.01 was used.

3.3 Evaluation Criteria

To assess Recovery Performance, *type I* and *type II error rate* of probing a qualitative treatment-subgroup interaction, *Cohen's κ agreement* and *splitting accuracy* were

[2]Bakeman proposed to apply the same guidelines of Cohen's f^2 to η^2_G. That is, to consider the size of 0.02 as small, 0.13 as medium and 0.26 as large [1].

evaluated.

Despite that there was no hypothesis test being conducted, the terms of "type I error rate" and "type II error rate" were borrowed to refer to the probability of wrongly detecting and failing to detect a qualitative interaction respectively, under the hypothesis stated that, "There is no qualitative treatment-subgroup interaction". They were utilized to inspect to what extend the methods can detect the interaction. Model 2 that represents the scenario with no qualitative interaction was applied for the evaluation of *type I error rate*. Whereas *type II error rate* was evaluated on data sets generated from Model 1, which suggests a scenario with the presence of qualitative interaction. The qualitative interaction is said to be detected when both partition subgroups P_1 and P_2 are present in the terminal leaves of solution trees. For the method of QUINT, this always holds true whenever a tree grows because of the *nonempty partition class condition*[3] it has to meet in its partitioning algorithm. Contrarily, a qualitative interaction is not ensured when a tree is grown by the method of PostZhang. For example, when one of the two terminal leaves that was originally classified to P_1 and P_2 by the method of Zhang has the leaf effect size smaller than the predefined *threshold.d*, the leaf is eventually assigned to the indifference group P_3 in PostZhang and thus no qualitative interaction is indicated in the PostZhang solution tree.

Cohen's κ agreement provides an alternative perspective of subgroups recovery performance. It measures the agreement between the assignment estimated by the methods and the true assignment. In addition to the common way of measuring assignment accuracy (i.e., the proportion of patients that are correctly assigned), *Cohen's κ* adapts the measurement with the probability of getting random agreement being considered.

Splitting accuracy measures the recovery performance with regard to the structure of derived trees. This can further be divided into three aspects, tree sizes (number of splits), splitting variables and split points. Firstly, size accuracy rate and the accuracy rate with bias of one being allowed, denoted as $P(\textit{True Size})$ and $P(\textit{True Size } \pm 1)$ accordingly, are inspected. They quantified the proportion of derived trees that recover the true tree size without and with a bias of size one being allowed.

Secondly, the proportion of capturing the correct splitting variables is inspected. The proportion is computed under certain conditions such that the ability of detecting true tree structure in each split is on focus. That is, when inspecting the first splitting variable, the proportion is computed under the condition of grown trees (i.e., trees with not only a root node). When inspecting the second splitting variable, it is computed conditionally on the first variable being detected correctly (regardless of the split

[3]The *nonempty partition class condition* in the partitioning algorithm of QUINT guarantees that the subsequent leaves after the first split are assigned to P_1 and P_2. If the condition is not met, no tree will be grown.

points). As the true tree is referred to Model 1 defined in subsection 3.1, which indicates the two splits on -0.433 of $X1$ and -0.219 of $X2$, the proportion of recovering the first and the second splitting variable are denoted separately as *P(Detect X1)* and *P(Detect X2)*.

Last but not least, among those splits which perform on the correct variables, the accuracy rate of splitting at correct split points rounded to one decimal place is measured, denoted as $P($*Split on the true split point of X1 |Detect X1*$)$ and *P(Split on the true split point of X2 |Detect X2)*, respectively to the first and the second split.

In addition to all the evaluation criteria measuring Recovery Performance, Predictive Performance is quantified by the absolute difference (AD) between the true expected treatment outcome and the estimated expected outcomes of derived trees, which is measured by the average treatment outcomes predicted by the trees. The true expected treatment outcome is equal to one (i.e., $E(Y) = 1$), according to the true model presented in (5) (See subsection 3.1). It represents the average outcome yielded under all patients being assigned to the true subgroups. The ADs are computed on the sample data as well as a test data. These are further referred to *sample AD* and *population AD*. When predicting the sample treatment outcomes, instead of taking the leaf optimal treatment mean as the predicted value, the outcomes of patients who are assigned to the indifference group are predicted by the average outcome of the belonging leaf, because of the fact that the indifference group suggests no preferable treatment. Intuitively, the *sample AD* are the difference between the true expected outcome, one, and the prediction made on the sample data that built the trees. However, to test whether the tree solutions (the assignment rules) work not only locally but also globally, the assignment rules derived from the sample trees are applied on a population data with size of 1,000,000 manipulated from Model 1. The average predicted outcomes with comparison to the true expected treatment outcome are referred to *population AD*.

Each of the evaluation criteria was obtained on the 2,400 data sets of the subjected model, namely Model 1 for all the evaluation criteria except for *type I error rate*, which was examined on the data sets generated from Model 2.

4 Results

4.1 Specification of the method tuning parameters

As different tuning parameters settings can clearly vary the results, we first searched for the optimal specification of the tuning parameters within each method, that is *dmin* parameter in method of QUINT and *threshold.d* parameter in method of PostZhang, before comparing the two methods.

4.1.1 QUINT: specification for *dmin*

The ANOVA analysis (Details are outlined in Appendix A.) on type I error rates with design factors, N, J, ρ and d, as the between subject factors and the *dmin* specification as the within subject factor revealed that type I error rates were largely affected by sample size N ($\eta^2_G > 0.02$) but not by the specification of *dmin* or other factors. A larger sample size generally induced a lower type I error rate regardless of the specification of *dmin*. Since *dmin*=0.4 marginally yielded lower type I error rates on both $N = 300$ and $N = 1000$, we suppose that the value of 0.4 could work well on our simulated data sets in terms of type I error rates. However, the ANOVA analysis of type II error rates showed that there was an interaction effect between N and d ($\eta^2_G = 0.038$). As shown in Figure 3, although type II error rates were satisfactory (< 0.2) in most of the situation, high type II error rates (> 0.5) emerged under small treatment effect ($d = 0.5$) along with large sample size ($N = 1000$), especially in the condition of *dmin*=0.4. Such high type II error rates were obviously undesirable. Thus in order to prevent from this unwanted situation, we proposed the specification of *dmin=0.4* for the simulated data sets with sample size $N = 300$ and *dmin=0.3* for the simulated data sets with sample size $N = 1000$, which is accordance with the recommendation by Dusseldorp et al. Due to the fact that treatment effects are normally unknown beforehand in the real world, the specification was only made for the sample size (N) but not for the treatment effect (d).

The Cohen's κ agreement was shown to be affected by the main effect of N and d according to its ANOVA analysis as shown in Appendix A. The larger the treatment effect and the sample size, the higher the agreement was. No remarkable difference was made when applying different values of *dmin*. (See details in Table 1). This implied that the earlier proposition of *dmin=0.3* for $N = 1000$ and *dmin=0.4* for $N = 300$ can be made without diminishing the agreement.

Table 1: The average Cohen's κ agreements of the simulated data sets with regards to different sample sizes (N=300, N=1000) and different design treatment effect sizes (d=0.2, 0.5 and 2) with different values of *dmin* parameter (0.3 and 0.4) being applied in the method of QUINT.

dmin	treatment effect size (d)	$N = 300$	$N = 1000$
0.3	0.5	0.22	0.26
	1	0.53	0.93
	2	0.92	0.99
0.4	0.5	0.17	0.12
	1	0.49	0.92
	2	0.92	0.99

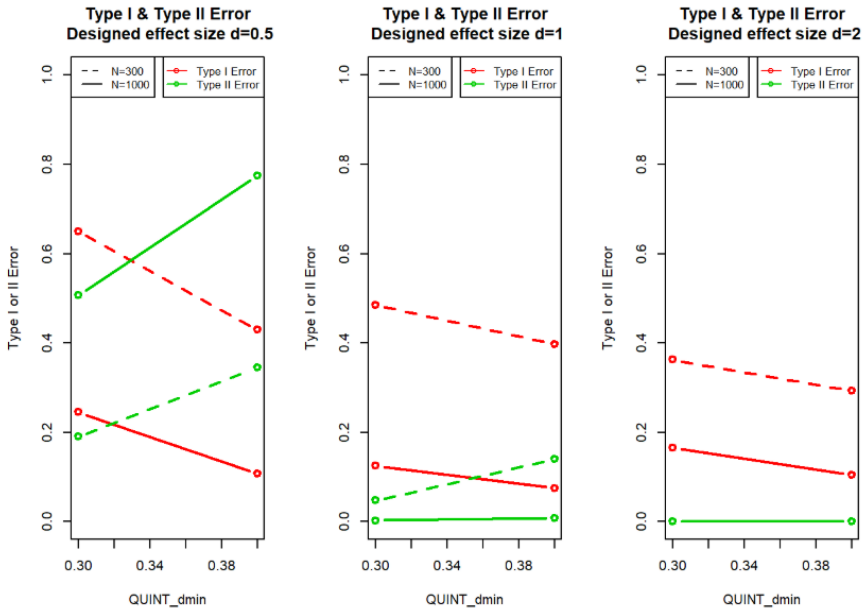


Figure 3: The plots display the type I error and type II error rate for method of QUINT with *dmin=0.3* and *dmin=0.4* with respect to the different treatment effect sizes and sample sizes. The panels from left to right draw the data set with medium (d=0.5), large (d=1), and extra large (d=2) designed effect size, respectively. The red lines represent the type I error rate while the green lines represent the type II error rate. In addition, solid lines and dashed lines respectively demonstrate the data sets with N=300 and N=1000.

4.1.2 PostZhang: specification for *threshold.d*

According to the ANOVA analysis, treatment effect size d slightly altered the type I error rates. The chance of committing a type I error was smaller when the treatment effect underlying in the data was smaller. None of the factors other than d, such as the specification of *threshold.d* and N was an influential factor. Since the underlying treatment effect is seldom known, the specification of *threshold.d* was decided mainly based on the ANOVA result of type II error rates and the Cohen's κ agreement.

The ANOVA analysis of type II error rates pointed out that *threshold.d*, treatment effect size d, and the interaction effect between them were factors ($\eta^2_G > 0.02$) that affected the type II error rates. The green lines in Figure 4 outlined the type II error rates under different treatment effect sizes and *threshold.d*. Despite the fact that there was some variation along treatment effect sizes and *threshold.d*, type II error rates were satisfactory in most of the time (i.e., type II error rate < 0.2). Only when treatment effects were small, setting a threshold above 0.3 suffered from a high risk of committing type II error. To have a conservative specification which could

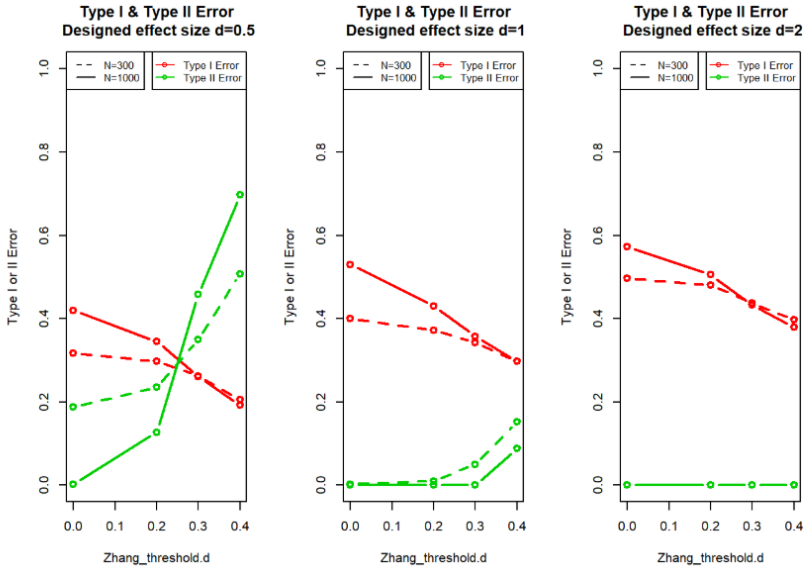


Figure 4: The plots display the type I error and type II error rate for method of PostZhang with *threshold.d* =0, 0.2, 0.3 and 0.4 with respect to the different treatment effect sizes. The panels from left to right draw the data set with d=0.5, 1, and 2, respectively. The red lines represent type I error rate while the green lines represent the type II error rate.

make the method sensitive to qualitative treatment-subgroup interaction even when the underlying treatment effect sizes were small, a threshold of 0.2 was preferred.

Regarding the Cohen's κ agreement, it was said to differ by factors N, d, ρ and the interaction between N and d according to the ANOVA analysis. Generally speaking, a lower correlation between the characteristics and a larger treatment effect size enhanced the agreement. A larger sample size also contributed to a higher κ agreement, especially when d was small. Yet, the value of *threshold.d* did not appear to influence the agreement. To sum up, to keep our simulation study sensitive to qualitative treatment-subgroup interaction even when the underlying treatment effect was small, we remained to set the tuning parameter *threshold.d* at 0.2 for the method of PostZhang.

4.2 Comparison between QUINT and PostZhang

With the analysis made within each method, the tuning parameters for the two methods were determined as follows:

- QUINT:
  dmin=0.4, for $N = 300$

dmin=0.3, for $N = 1000$

- PostZhang:
  threshold.d=0.2

Using these tuning parameters, the method of QUINT and the method of PostZhang were compared regarding their Recovery Performance and Predictive Performance. The variables that yielded η^2_G larger than 0.02 in the ANOVAs are shown in Table 3.

Table 2: The average Cohen's κ agreements of the simulated data sets with regards to different sample sizes (N=300, N=1000) and different design treatment effect sizes (d=0.2, 0.5 and 2) with different values of *threshold.d* parameter (0, 0.2, 0.3 and 0.4) being applied in the method of PostZhang.

threshold.d	*d*	N=300	N=1000
0.0	0.5	0.27	0.49
	1	0.50	0.53
	2	0.53	0.53
0.2	0.5	0.28	0.52
	1	0.51	0.57
	2	0.56	0.60
0.3	0.5	0.27	0.52
	1	0.52	0.62
	2	0.58	0.64
0.4	0.5	0.24	0.45
	1	0.54	0.65
	2	0.61	0.68

4.2.1 Type I /II Error Rate

The ANOVA analysis in Table 3 showed that the type I error rates did not differ between the methods of QUINT and PostZhang, nor did it alter on the other factors. Meanwhile, the type II error rates were mainly influenced by factors *method*, *treatment effect*, the interaction effect between them, and the interaction effect between *sample size* and *treatment effect*. As seen in Table 4, the smaller the treatment effect sizes, the higher the chance to commit a type II error. In addition while enlarging the sample size could normally reduce type II error rates, it seemed not the case under small treatment effect size. It was particularly the case for the method of QUINT. When the underlying treatment effect size was small (e.g., $d = 0.5$), rather high type II error rates appeared by the method. Except for that, type II error rates were satisfactory (< 0.20) in most of the situations.

4.2.2 Cohen's κ Agreement

According to the ANOVA results of Cohen's κ agreement shown in Table 3, *method*, *sample size*, *treatment effect*, the two-way interaction between *method* and *treatment*

Table 3: The η^2_G obtained via ANOVA on each evaluation. Only the variables that have η^2_G above 0.02 were recorded in the table. Method is a variable with two categories: *QUINT* and *PostZhang*. N, d are the design factors specified in section 3. ":" signified the interaction term between factors. The variables without any $\eta^2_G > 0.02$ are neglected in the table.

	Method	N	d	Method:N	Method:d	N:d
Recovery Performance						
Type I error rates	-	-	-	-	-	-
Type II error rates	0.032	-	0.216	-	0.026	0.038
Cohen's κ Agreement	0.088	0.141	0.451	-	0.242	0.023
P(True Size)*	0.303	0.03	0.169	0.044	0.193	
P(True Size $\pm$ 1)*	-	-	0.104	-	0.059	0.021
Predictive Performance						
Sample AD	-	0.122	0.045	-	-	-
Population AD	0.088	0.099	0.155	-	-	-

* Among the evaluations of splitting accuracy, only the evaluation relevant to recovering the true tree size, that is, P(True Size and P(True Size $\pm$ 1), were analyzed using ANOVA. The other measurements concerning the splitting variables and split points were inspected in an exploratory perspective.

effect, and the two-way interaction between *sample size* and *treatment effect* had remarkable effects on the performance of the subgroup assignment agreement. Generally speaking, the larger the sample size and the treatment effect, the easier for the methods to recognize the underlying structure and thus yielded a higher assignment agreement to the true scenario. Among them, the method of QUINT had an overall better assessment compared to PostZhang. When the treatment effect was large (e.g., $d = 2$), QUINT could reach a very good ($\kappa > 0.8$) agreement to the true scenario as shown in Table 4. However, in the condition of small treatment effect, the agreements QUINT reached were lower than those assessed by PostZhang.

To explore the possible reason that induced the relative low agreement of the method of PostZhang, we further inspected the proportion of patients in each subgroup that were correctly classified. Figure 5 showed that the inefficient performance in recognizing the indifference group was apparently the main reason. It revealed the fact that the post-hoc indifference group assignment was an approach that could only identify the indifference group conservatively.

4.2.3 Splitting Accuracy

Suggested by the ANOVA results, *method*, *sample size*, *treatment effect*, interaction between *method* and *treatment effect*, and the interaction between *method* and *sample size* were influential factors against the size accuracy *P(True size)*. As seen in the solid lines of Figure 6, the method of QUINT significantly outperformed the method of PostZhang on capturing the exact complexity (size) of the true tree, particularly

Table 4: The average evaluation outcomes computed marginally on *sample size* N and *treatment effect* d. The evaluation criteria: type I error rates, type II error rates and Cohen's κ agreement are displayed.

Method	**QUINT**			**PostZhang**		
Type I error rates						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.430	0.398	0.293	0.298	0.373	0.480
N=1000	0.245	0.125	0.165	0.345	0.430	0.505
Type II error rates						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.345	0.140	0.00	0.235	0.01	0.00
N=1000	0.507	0.002	0.00	0.128	0.00	0.00
Cohen's κ agreement						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.17	0.487	0.923	0.276	0.509	0.556
N=1000	0.26	0.927	0.986	0.521	0.574	0.596

in the condition of the large treatment effect size. In the method of QUINT, the size accuracy increased along with the sample size and treatment effect size. When the treatment effect size was medium (i.e., $d = 1$), the size accuracy was improved remarkably by adding sample size from $N = 300$ to $N = 1000$. Yet, it was not the case for the method of PostZhang. However, the method of PostZhang could derive the true tree size with a bias of one (plotted with the dotted lines in Figure 6) being allowed most of the times. The method of PostZhang even performed better than QUINT under the small treatment effect size, in terms of the accuracy rate for capturing the true tree size with a bias of one being allowed.

The further exploration on splitting variables and split points were shown in Table 5. Regarding the first split, both methods have impressive performance on recognizing the correct variable, $X1$, when a tree grew. Even when the sample size and treatment effect size were small (e.g., $N = 300$, $d = 0.5$), $X1$ could be recognized with the probability around 0.7. In addition, except for the extreme situation when both the treatment effect and the sample size were small, both methods nicely perform the split on a split point which was approximate to that of the true model.

The second splitting variable in the true model (i.e., $X2$) was successfully detected at the second split most of the situations under $X1$ was recognized in the previous split in both methods. However, again when under the extreme situation, that is, small

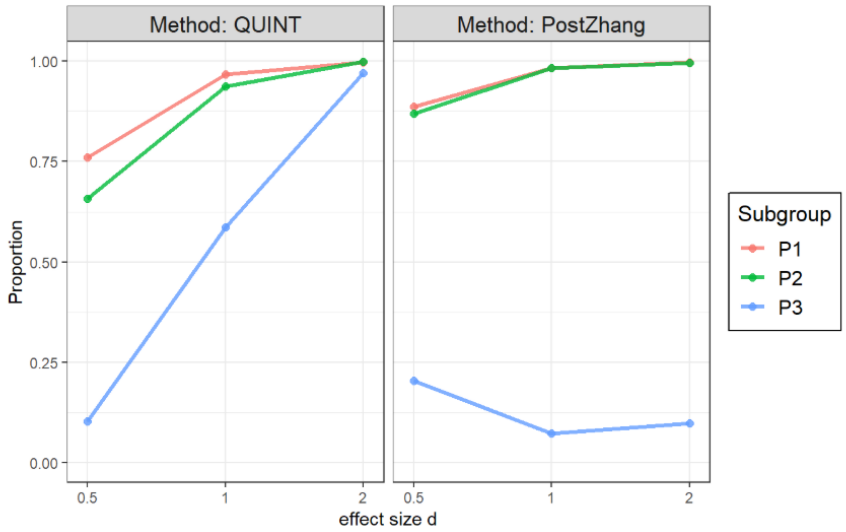


Figure 5: the proportion of patients in each subgroup that are correctly assigned in method of QUINT and method of PostZhang. The red, green and blue lines separately represent the measurements in subgroup P_1, P_2 and P_3.

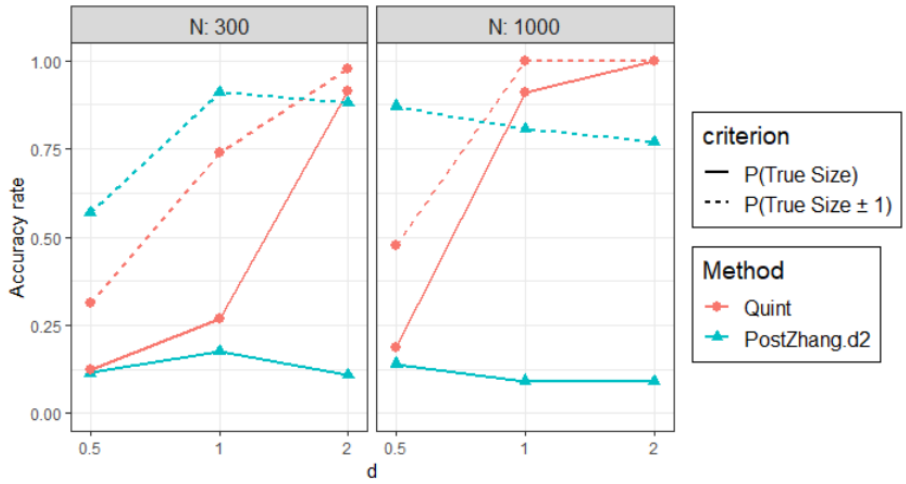


Figure 6: The Splitting size accuracy with and without bias of one, with regards to different treatment effect sizes (d) and sample sizes (N), for the method of QUINT and PostZhang. The solid line represents the proportion of capturing the exact size of the true model, *P(True size)*. The dotted line represents the proportion of capturing the true size plus or minus one, *P(True size* ± 1*)* .

Table 5: The evaluation on splitting accuracy criterion with respect to the splitting variables and split points. The proportion of recovering the first and the second splitting variables (X1 and X2) are denoted as *P(Detect X1)* and *P(Detect X2)*, respectively. They were measured under the conditions mentioned in subsection 3.3. Under the splitting variables being detected, the chance of these splits performing on the same split point as in the true model (rounded to the first decimal place) is denoted as $P(Split\ true\ X1\ point|Detect\ X1)$ and $P(Split\ true\ X2\ point|Detect\ X2)$, accordingly. The evaluation were derived from Model 1.

Method	**QUINT**			**PostZhang**		
P(Detect X1) ($P(Split\ on\ the\ true\ split\ point\ of\ X1|Detect\ X1)$)						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.65(0.26)	0.98(0.59)	1.00(0.80)	0.70(0.28)	0.97(0.61)	1.00(0.87)
N=1000	0.99(0.59)	1.00(0.80)	1.00(0.94)	1.00(0.55)	1.00(0.86)	1.00(0.97)
P(Detect X2) ($P(Split\ on\ the\ true\ split\ point\ of\ X2|Detect\ X2)$)						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.32(0.20)	0.85(0.32)	1.00(0.53)	0.32(0.20)	0.88(0.29)	0.99(0.51)
N=1000	0.85(0.30)	1.00(0.59)	1.00(0.84)	0.64(0.35)	0.98(0.61)	1.00(0.85)

sample size and small treatment effect size, it was harder for the methods to detect X_2 (i.e., chance around 0.32). Nevertheless, such issue could be overcome largely by adding more sample size, especially for the method of QUINT. Among the schemes when $X2$ was recognized, the split points could be accurately observed up to one decimal place when the treatment effect and the sample size were both sufficiently large (e.g., $d \geq 1$, $N = 1000$).

4.2.4 Sample and Population Absolute Difference to the True Expected Outcome

As shown in Table 6, the distances between the tree expected outcomes and the true expected outcomes were all close to zero, which indicated that the predictive performance was satisfying in both methods. According to the ANOVA results (shown in Table 3), both sample AD and population AD decreased along with increasing sample sizes and treatment effect sizes. Apparently, when trees were trained using data with larger sample sizes and greater treatment effect sizes, it was easier to get the optimal solutions. While sample ADs were obtained without noticeable difference between the two methods, the population ADs were different between the two methods. Generally speaking, the population expected outcomes obtained via the method of PostZhang were closer to the true treatment outcome, which is equal to one.

Table 6: The average evaluation outcomes computed marginally on *sample size N* and *treatment effect d*. The evaluation criteria: Sample Absolute Difference to the true treatment outcome (Sample AD) and Population Absolute Difference to the true treatment outcome (Population AD) in each method are displayed.

Method	**QUINT**			**PostZhang**		
Sample AD						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.114	0.101	0.078	0.120	0.088	0.085
N=1000	0.098	0.037	0.037	0.052	0.049	0.045
Population AD						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.123	0.092	0.044	0.090	0.024	0.011
N=1000	0.095	0.013	0.010	0.025	0.006	0.003

5 Discussion

The study was conducted to compare the two subgroup identification approaches, QUINT and the method of PostZhang in true scenarios including an indifference group in terms of their Recovery Performance and Predictive Performance. The hypothesis which stated that QUINT had superior performance on recovering the underlying subgroups was confirmed under the circumstances that the treatment effect sizes were sufficient. With regard to the predictive performance, the method of PostZhang showed higher accuracy than that of the method of QUINT.

The first finding that QUINT outperformed the method of PostZhang on recovering the subgroups in the true scenario was supported by the overall higher Cohen's κ agreement and the size accuracy rate achieved by QUINT. The assignment rules recovered by QUINT are more accurate with respect to their complexity (number of splits) and can classify the patients to the correct subgroups to a great extend. In contrast, more assignment rules with inaccurate complexity are derived by PostZhang, along with lower agreements of the subgroups assignment to the true scenario. They are believed to be subject to the partitioning algorithm of the methods, in which the indifference group is considered while growing trees by QUINT but not by PostZhang. Since PostZhang focus on purifying the classes (i.e., the direction of estimated contrast values) in nodes while growing trees, the algorithm tends to avoid a node behaves like an indifference group, for example, containing similar numbers of patients who benefit more from treatment A and who benefit more from treatment B. As a consequence, the complexity is seldom accurate. In addition, the post-hoc indifference group assignment of PostZhang makes it conservative when recognizing the indifference group. It

implies that many patients who belong to the indifference group will be classified to an alternative treatment group and leads to lower subgroup assignment agreements in comparison to that of the method of QUINT.

Nevertheless, the method of QUINT was also pointed out to have disadvantage when the treatment effect was small. Under the small treatment effect, PostZhang achieves higher agreement to the true subgroups assignment and the derived complexity is more precise in terms of the accuracy rate of obtaining tree sizes within one bias. This can link to the high type II error rates that QUINT suffers from when the underlying treatment effect in the data is small. When a type II error occurs (i.e., no qualitative treatment-subgroup interaction is detected) in QUINT, no tree is grown. Many of the derived solutions will thus have the tree size and the subgroup agreement equal to zero.

The second finding in our simulation study shows that the method of PostZhang can yield population expected outcomes that are closer to the true expected outcome, which is the expected outcome when all patients are assigned to the true subgroups. We found this results unsurprising due to the fact that it is originated from the method of Zhang et al., which aimed at seeking the regime that maximizes the estimate of the expected outcome. It is reasonable that the inefficient in post-hoc indifference group assignment of PostZhang does not affect the prediction of population expected outcomes, since those belonging to the indifference group are estimated with the same value regardless which group they are assigned to. The estimated population expected outcome of PostZhang remained identical to that derived from the original method of Zhang and thus have the same advantage of the predictive performance. A previous comparison study of Sies and van Mechelen (2017) has demonstrated that the method of Zhang et al. performed the best regarding the expected outcomes among various subgroup identification approaches including the method of QUINT.

Our simulation study demonstrated the recovery performance and predictive performance of QUINT and PostZhang under the true scenario including an indifference group. The conclusion is made under a simulated scenario which pertains a rather simple structure. It is doubtful whether the conclusion is applicable on a more complex scenario, for instance, three or more splits in the true model. The simulation study conducted by Sies and van Mechelen (2017) has results stated that less proportion of patients are classified correctly in the method of QUINT compares to that of the method of Zhang et al. under both simple and complex scenarios. However, the study was conducted under no indifference group being considered. Additional experiments are required to inspect the recovery performance in complex scenarios with the indifference group being included.

Also, the choice of threshold for the post-hoc indifference group assignment in PostZhang alters the type II error rates. If one would have some pre-knowledge about the expected treatment effect size of their study and chose an appropriate threshold

based on that, the recovery performance of PostZhang may be improved and resolve the conservative nature of the method to some extend. Yet, we do not expect the improvement to be remarkable since it is shown in our study that the agreement did not differ between the choices of the threshold.

To sum up, although PostZhang is less efficient in identifying the indifference group, it has advantage of deriving treatment outcomes which are more accurate. It is a practical issue to consider the indifference group. If one wants to put emphasize not only on the optimal treatment but also the identification of the indifference group, then QUINT can provide more insight on the possible characteristics that compose this group.

References

[1] Roger Bakeman. Recommended effect size statistics for repeated measures designs. *Behavior Research Methods*, 37(3):379–384, 2005.

[2] Jacob Cohen. *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Earlbaum Associates, New York, 1988.

[3] Elise Dusseldorp and Iven Van Mechelen. Qualitative interaction trees: a tool to identify qualitative treatment-subgroup interactions. *Statistics in medicine*, 33(2):219–237, 2014.

[4] Ilya Lipkovich, Alex Dmitrienko, and Ralph B. D'Agostino Sr. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in medicine*, 36(1):136–196, 2017.

[5] Donald B. Rubin. Bayesian inference for causal effects: The role of randomization. *The Annals of Statistics*, 6(1):34–58, 1978.

[6] Aniek Sies and Iven Van Mechelen. Comparing four methods for estimating tree-based treatment regimes. *Int. J. Biostat*, 13(1), 2017.

[7] Baqun Zhang, Anastasios A. Tsiatis, Marie Davidian, Min Zhang, and Eric Laber. Estimating optimal treatment regimes from a classification perspective. *Stat*, 1(1):103–114, 2012.

A

ANOVA within the two methods

The results of ANOVA within each methods are shown in the below table. The analyses were conducted using evaluation criteria relevant to the recovery performance as dependent variables. The different specifications of tuning parameters, that is *dmin* with levels 0.3 and 0.4 for the method of QUINT and *threshold.d* with levels 0, 0.2, 0.3 and 0.4 for the method of PostZhang, were applied as the within-subjects variables and design factors N, J, ρ and d were the between-subjects variables. In addition, the interaction effect of these variables were also considered. A variable was said to have leading influence on the dependent variables if its generalized-eta-squared (η^2_G) exceeds 0.02.

Table 7: The η^2_G obtained via ANOVA on each evaluation within the two methods, QUINT and PostZhang. The factor "Method" here are referred to the method with different specification of *dmin* and *threshold.d* for QUINT and PostZhang, respectively. Only the variables that have η^2_G above 0.02 are displayed in the table.

QUINT	Method	N	d	ρ	Method:d	N:d
Type I error rates	-	0.116	-	-	-	-
Type II error rates	-	0.022	0.216	-	-	0.038
Cohen's κ Agreement	-	0.099	0.603	-	-	-
PostZhang	-	-	-	-	-	-
Type I error rates	-	-	0.021	-	-	-
Type II error rates	0.078	-	0.152	-	0.067	-
Cohen's κ Agreement	-	0.116	0.193	0.030	-	0.044

B

Sample and Population Expected Outcome

Although the *sample ADs* revealed no significant difference between the method of QUINT and PostZhang, the differences appeared when we directly inspected the sample expected outcomes. The η^2_G of the factor "Method", which indicates the method implemented (i.e., QUINT or PostZhang), was greater than 0.02 when conducting ANOVA on sample and population expected outcomes. As shown in Table 8, the sample expected outcomes derived by PostZhang were higher than that by QUINT. However, it was possible to be caused by the conservative nature of PostZhang relating

the post-hoc indifference group assignment. Many patients who belong to the indifference group ended up being assigned to alternative treatment group and were estimated to gain more benefits than it actually did as the prediction was made using the optimal average treatment outcome in the assigned leaf for the sample expected outcome. Nevertheless, the prediction made in population was not based on the optimal average treatment outcomes in leaves, thus the concern did not violate the conclusion that PostZhang has better predictive performance.

Table 8: The average evaluation outcomes computed marginally on *sample size* N and *treatment effect* d. The evaluation criteria: Sample Expected Outcome and Population Expected Outcome in each method are displayed.

Method	**QUINT**			**PostZhang**		
Sample Expected Outcome						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	1.024	0.978	0.984	1.054	1.049	1.051
N=1000	0.941	1.003	0.998	1.017	1.032	1.026
Population Expected Outcome						
	d=0.5	d=1	d=2	d=0.5	d=1	d=2
N=300	0.877	0.908	0.956	0.910	0.976	0.989
N=1000	0.905	0.987	0.990	0.975	0.994	0.997