Received: 11 May 2018 | Revised: 15 February 2019 | Accepted: 22 April 2019

DOI: 10.1002/pst.1951

MAIN PAPER

WILEY

A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations

Cynthia Huber[1] | Norbert Benda[2,1] | Tim Friede[1]

[1]Department of Medical Statistics, University Medical Center Göttingen, Göttingen, Germany

[2]Federal Institute for Drugs and Medical Devices (BfArM) Research Department, Bonn, Germany

Correspondence
Cynthia Huber, Department of Medical Statistics, University Medical Center Göttingen, Göttingen 37075, Germany.
Email:
cynthia.huber@med.uni-goettingen.de

With advancement of technologies such as genomic sequencing, predictive biomarkers have become a useful tool for the development of personalized medicine. Predictive biomarkers can be used to select subsets of patients, which are most likely to benefit from a treatment. A number of approaches for subgroup identification were proposed over the last years. Although overviews of subgroup identification methods are available, systematic comparisons of their performance in simulation studies are rare. Interaction trees (IT), model-based recursive partitioning, subgroup identification based on differential effect, simultaneous threshold interaction modeling algorithm (STIMA), and adaptive refinement by directed peeling were proposed for subgroup identification. We compared these methods in a simulation study using a structured approach. In order to identify a target population for subsequent trials, a selection of the identified subgroups is needed. Therefore, we propose a subgroup criterion leading to a target subgroup consisting of the identified subgroups with an estimated treatment difference no less than a pre-specified threshold. In our simulation study, we evaluated these methods by considering measures for binary classification, like sensitivity and specificity. In settings with large effects or huge sample sizes, most methods perform well. For more realistic settings in drug development involving data from a single trial only, however, none of the methods seems suitable for selecting a target population. Using the subgroup criterion as alternative to the proposed pruning procedures, STIMA and IT can improve their performance in some settings. The methods and the subgroup criterion are illustrated by an application in amyotrophic lateral sclerosis.

KEYWORDS

decision trees, predictive biomarker, personalized medicine, treatment-by-subgroup interactions

1 | INTRODUCTION

With the advances in technologies such as genomic sequencing biomarkers have become useful tools in drug development. To date a number of drugs have been authorised in biomarker-defined subgroups.[1] Usually, such a stratification

The views expressed in this article are the authors personal views and do not necessarily represent those of the Federal Institute for Drugs and Medical Devices (BfArM).

is based on a predictive biomarker implying a treatment-by-biomarker interaction and determining the effect of a therapeutic intervention. A predictive biomarker has to be distinguished from a prognostic one, which predicts the course of a disease. In our context, the term biomarker may not just be a genetic marker but could also refer to other baseline patient characteristics such as demographic variables.[2]

In many cases, the stratification of patients is purely based on the prior knowledge of the drug's mechanism of action. If, however, the mechanism of a drug is not fully understood, data-driven evidence of a treatment-by-subgroup interaction can support the pharmacological and biological reasoning of a biomarker's predictivity. This was the case for the drugs panitumumab and cetuximab with the KRAS marker. Retrospective analyses showed that only wild-type KRAS patients benefit from these treatments. Based on this finding, the inhibition of the RAS/RAF/MAPK pathway was considered to be responsible for the activity of panitumumab and cetuximab, two anti-epidermal grow factor receptor (EGFR) agents. Since the retrospective analyses were convincing, further studies were conducted only in biomaker-selected (ie, wild-type KRAS) patients.[3-7]

For the drugs panitumumab and cetuximab differential treatment effects in subgroups were hypothesized upfront. If, however, there is no prior hypothesis regarding a subgroup with an enhanced treatment effect available, exploratory subgroup identification methods, which will be considered below, are frequently applied in order to find predictive biomarkers and generate such hypotheses. A possibility of incorporating findings regarding potential treatment-by-subgroup interactions in subsequent trials is by selecting the study design accordingly, eg, by multi-population designs or adaptive enrichment designs.[8-12] The interest in methods for identifying subgroups increased over the last years. Ondra et al[13] identified 86 articles on the topic of identification and confirmation of targeted subgroups in a systematic review of the literature. The tutorial on subgroup identification by Lipkovich et al[14] mentions around 60 articles related to subgroup identification methods. Usually, they were developed for one of the two following frameworks for personalized medicine. The first framework aims at identifying patients for a given treatment and is therefore related to a search for quantitative treatment-by-biomarker interactions, whereas the second framework aims at finding the right treatment for a given patient resulting in a special interest in qualitative treatment-by-biomarker interactions. In the presence of a qualitative interaction, different patients benefit from different treatments. A subset of patients with specific biomarker values will profit from the experimental treatment, whereas the complementary subgroup will not benefit or will even be harmed by the experimental treatment compared with the control. For the first framework, several tree-based methods were proposed. These have the advantage of identifying predictive biomarkers and selecting cut-off values in case of continuous predictive biomarkers. The selection of cut-off values for continuous predictive biomarkers is needed in order to get decision rules for subgroups. Interaction trees (IT),[15] model-based recursive partitioning (MOB),[16] subgroup identification based on differential effect search (SIDES),[17] and simultaneous threshold interaction modeling algorithm (STIMA)[18] are examples for methods with the aim of identifying subgroups with an enhanced treatment effect. The adaptive refinement by directed peeling algorithm (ARDP) for subgroup identification as included in Patel et al[19] is also an example for a subgroup identification method identifying the predictive biomarker with its corresponding cut-off value.

Although there are numerous methods for this purpose available, comparisons of methods applicable to similar settings[20] are lacking, with the notable exceptions of articles by Doove et al,[21] Alemayehu et al,[22] and Sies and Mechelen.[23] Boulesteix et al[20] point out that more neutral comparison studies, which evaluate the behavior of existing methods, are needed. Also, the existing comparisons focus on emphasizing differences between methods and thereby not always consider scenarios relevant for drug development.

In drug development, it is often of interest to identify one subgroup with a compelling treatment effect for defining an enrichment strategy for future studies and to refine the envisaged indication. Most of the subgroup identification methods identify multiple subgroups. Therefore, those subgroups have to be selected that should form the target population. We propose a subgroup criterion for this selection. The treatment effects in each of the identified subgroups have to exceed a pre-specified threshold in order to be assigned to a potential future target population, the biomarker-positive $BM+$ subgroup. Note that the treatment effect in the BM+ subgroup is not necessarily the average over the subgroups' treatment effects in all settings due to usually different sizes of the identified subgroups. The effect in the target population may be smaller than the threshold in case the realized treatment allocation ratios differ in the identified subgroups, which is, however, in general, expected to be negligible in randomized trials.

Here, we present the methods IT, MOB, STIMA, SIDES, and ARDP in consistent notation, which makes similarities and differences more apparent. Furthermore, we compare their operation characteristics performances for the selection of a target population in a Monte-Carlo simulation study using a structured approach.[24] Selecting a target population has the same aim as companion diagnostics. Both aim at identifying "patients who are most likely to benefit from the corresponding medicinal product."[25] Therefore, we use performance measures like accuracy, sensitivity, and specificity.[26] Furthermore, we evaluate the Type I and Type II error rate for these methods applying the proposed subgroup criterion. Since we are interested how the performance is influenced by certain data characteristics, we consider different sample sizes, effect sizes, and data-generating mechanism influencing the true target subgroup, the *BM*+ subgroup.

The remainder of this paper is organized as follows. In Section 2, we briefly outline the five methods. In the following section, we describe how subgroup definitions can be obtained by the results of the five considered methods, and we introduce the subgroup criterion used for selecting the target population. In Section 4, we illustrate the methods and the subgroup criterion by an amyotrophic lateral sclerosis (ALS), and in Section 5, we will describe the simulation study and present its results. The manuscript concludes with a discussion.

2 | METHODS FOR SUBGROUP IDENTIFICATION

In this section, we will introduce the notation, and we will briefly describe the five methods for subgroup identification we compared in a simulation study. Since the range of methods proposed for subgroup identification is wide, we focused on methods, which are able to select cut-off values in the presence of continuous biomarkers and were developed for continuous outcome variables. Tree-based methods as IT,[15] MOB,[16] SIDES[17], and STIMA[18] fulfill these requirements. The adaptive refinement by directed peeling algorithm (ARDP) for subgroup identification proposed by Patel et al[19] uses a peeling procedure, which identifies a peeling variable and a cut-off value in every iteration step. This is comparable with the four recursive partitioning methods and can also be described with tree analogies.

We will only consider the situation of a randomized controlled clinical trial. Patients included in the trial receive either an experimental treatment, denoted $T = 1$, or a control treatment, denoted $T = 0$. Besides of the outcome variable Y, the data include p candidate biomarkers for each of the n patients denoted by $\mathbf{X} = X_1, \dots, X_p$. The observed data $(Y_i, T_i, \mathbf{X})$ for patient i, with $i = 1, \dots, n$, are assumed to be independent and identically distributed across i. Based on the different handling of the the selection bias and the availability of many biomarkers measured on a continuous scale, we consider only continuous biomarkers. In order to define a subgroup, those biomarkers have to be dichotomized. Without loss of generality, we assume that larger values of the outcome are preferable.

We denote an identified subgroups by $\hat{S}$. The expected outcomes in the identified subgroups for patients in the control and the experimental treatment arm are denoted by $\mu_0(\hat{S}) = E(Y|T = 0, \mathbf{X} \in \hat{S})$ and $\mu_1(\hat{S}) = E(Y|T = 1, \mathbf{X} \in \hat{S})$, respectively.

2.1 | Interaction trees

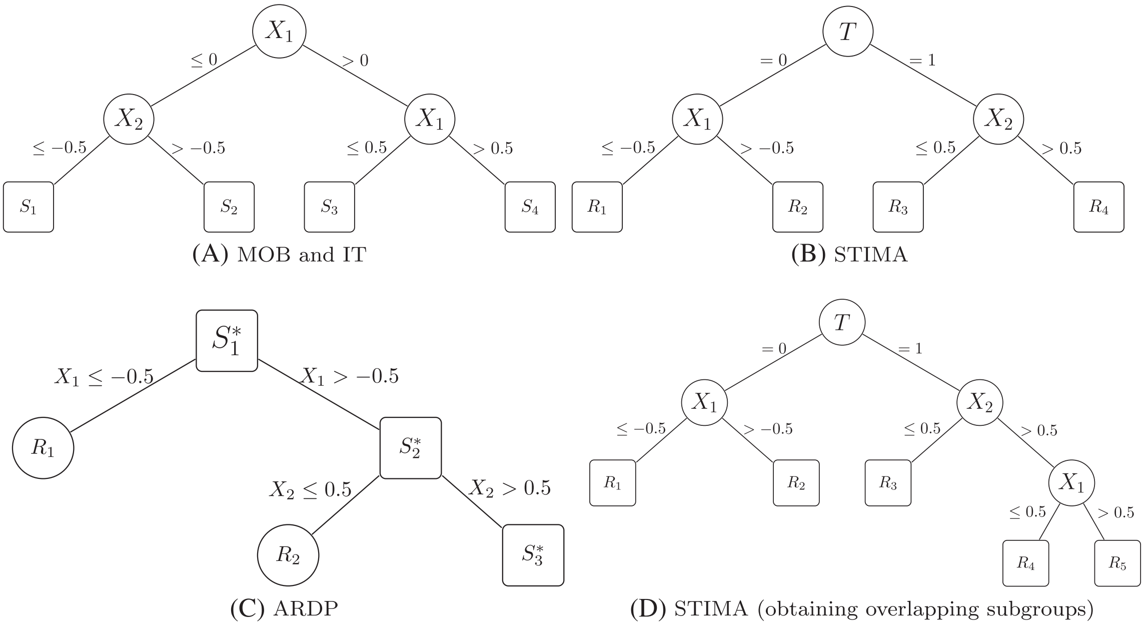
Interaction trees were developed for exploring the heterogeneity of treatment effects. Su et al[15] use the CART methodology[27] for tree construction, which consists of three steps: growing a large initial tree, pruning, and selecting the best-sized pruned tree.

In order to construct the tree, we associate each node, root, and child node, with the following linear regression model:

$$E(Y|\mathbf{X}) = \alpha + \beta_0 \cdot T + \gamma \cdot I(X_j \leq c) + \beta_1 \cdot T \cdot I(X_j \leq c) \quad \text{with} \quad j = 1, \dots, p.$$

As splitting criterion, the squared t test for testing the hypotheses $H_0 : \beta_1 = 0$ in the above mentioned model is used. The model includes the main effect of the treatment indicator variable and of the biomarker considered for the split. Biomarkers not involved in the split are not included as main effects in the model. The split associated with variable X_j and the split point c yielding the maximum t^2-statistic is used for the split. The splitting is repeated recursively in the obtained child nodes until a stopping criterion (eg, minimum number of observations in a node, node is pure) is met.

This procedure leads to a large initial tree, which is then pruned. The pruning procedure described in Breiman et al[27] was used. This pruning procedure results in a sequence of nested subtrees by iteratively truncating the "weakest link" of the tree. Afterwards, a single subtree is chosen from the sequence as final tree. For both determining the "weakest link" and the best-sized subtree selection, an interaction-complexity criterion was developed by Su et al, which is analogous to the split complexity measure proposed by LeBlanc and Crowley.[28] Since the interaction-complexity measure is used for truncating and the selection of the best-sized tree, the selection might be over optimistic. In order to get an unbiased estimate of the interaction-complexity measure for determining the best-sized subtree, an independent test set or a bootstrap procedure should be used. Figure 1 shows how a tree obtained by IT can look like.



(A) MOB and IT

(B) STIMA

(C) ARDP

(D) STIMA (obtaining overlapping subgroups)

FIGURE 1 Theoretical example of trees obtained by the different methods

2.2 | Model-based recursive partitioning

MOB for subgroup analyses proposed by Seibold et al[16] detects parameter instabilities in the treatment effect of the overall population. The trees obtained by MOB look similar to the ones obtained by IT. But they differ in their growing procedure. Like IT, MOB associates every node of the tree with a parametric model denoted by $\mathcal{M}((Y,T,X),\theta)$. The vector $\boldsymbol{\theta} = (\boldsymbol{\alpha},\boldsymbol{\beta},\boldsymbol{\gamma},\boldsymbol{\sigma})$ denotes the parameter vector fitting the data $(Y,T,\mathbf{X})$. The intercepts are denoted by $\boldsymbol{\alpha}$, $\boldsymbol{\beta}$ denotes the treatment effects, $\boldsymbol{\gamma}$ refers to other interesting effects, and $\boldsymbol{\sigma}$ to the nuisance parameters. The parameter vector can be estimated by minimizing an objective function $\Psi((Y,X),\theta)$ (eg, the negative log-likelihood). We use the linear regression model $E(Y|\mathbf{X},T) = \alpha + \beta_0 \cdot T$ for every node.

The idea underlying MOB is that in the presence of subgroups, there is not a single global model fitting the data well. Therefore, MOB splits the data when intercept or treatment effect differ across subgroups. Thus, MOB aims at detecting parameter instabilities. Since a correlation between the partial score function of $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$, ie, $\psi_\alpha((Y,T,\mathbf{X}),\boldsymbol{\theta}) = d\Psi((Y,T,\mathbf{X}),\boldsymbol{\theta})/d\boldsymbol{\alpha}$ and $\psi_\beta((Y,\mathbf{X}),\boldsymbol{\theta}) = d\Psi((Y,T,\mathbf{X}),\boldsymbol{\theta})/d\boldsymbol{\beta}$, and the covariates corresponds to parameter instabilities, MOB tests for independence using the M-fluctuation test introduced by Zeileis et al.[29] The tested hypotheses are as follows:

$$H_0^{\alpha,j} : \psi_\alpha((Y,\mathbf{X},T),\hat{\boldsymbol{\theta}}) \perp X_j, \quad j = 1, \ldots, p$$

$$H_0^{\beta,j} : \psi_\beta((Y,\mathbf{X},T),\hat{\boldsymbol{\theta}}) \perp X_j, \quad j = 1, \ldots, p.$$

A partition is only performed if at least one of the $2 \times p$ null hypotheses can be rejected at a pre-specified nominal level. For this splitting procedure, MOB uses multiplicity adjustments. The partitioning variable X_{j*} is associated with the maximum correlation to any of the partial score functions. In order to obtain the cut-off value for the chosen variable X_{j*}, we have to sum the objective functions of the conceivable subsets. We get the cut-off value by optimizing this segmented objective function. This procedure is recursively applied until we cannot reject any of the independence hypotheses. Since MOB applies this pre-pruning procedure, the obtained final tree is the best-sized tree, and no further pruning and selection steps are necessary.

2.3 | Simultaneous threshold interaction modeling algorithm

STIMA proposed by Dusseldorp et al[18] was developed for overcoming the drawbacks of both additive models in the presence of higher order interactions and tree-based methods like CART[27] in the presence of linear main effects. Therefore, STIMA can also be used to detect interactions with the treatment variable. The Simultaneous Threshold Interaction Modeling Algorithm associates the whole tree with a linear model. STIMA uses a linear regression model for modeling the main effects and a tree for the higher-order interactions. Both, the main effects and the threshold interaction effects, are optimized simultaneously. Since STIMA combines a parametric and nonparametric approach, it is called a hybrid model. Although STIMA was developed for a much broader purpose, we use it here for subgroup identification.

For subgroup identification purposes, it is necessary to force the procedure to make the first split at the treatment indicator variable. Furthermore, we need to exclude the treatment variable from the linear main effects in order to avoid linear dependencies. Therefore, the initial reference model is as follows:

$$E(Y|\mathbf{X}) = \alpha + \beta_0 I(T = 1) + \sum_{j=1}^{p} \gamma_j X_j. \tag{1}$$

STIMA performs an exhaustive search among all covariates and all splits for each of the two child nodes of the root (childnodes $T = 0$ and $T = 1$) in order to find further splits defined by a split variable X_{j*} and a split point c^*. STIMA uses the realtive increase in variance accounted for by an expanded model. This criterion is defined as follows: $f_l^2 = (R_l^2 - R_{l-1}^2)/(1 - R_l^2)$ with R_l^2 as the coefficient of determination estimated before and after split l. R_l^2 after split l is defined as $R_l^2 = \sum_i (\hat{Y}_{il} - \bar{Y})^2 / \sum_i (Y_i - \bar{Y})^2$, where $\hat{Y}_{il}$ refers to the predicted value for patient i with the model induced by split l, and $\bar{Y}$ denotes the mean of the observed data. Variable X_{j*} and split point c^* induce the the highest increase in f_l^2.

For the first split at node $T = 1$, the current reference model (Equation 1) is compared with the following expanded model:

$$E(Y|\mathbf{X}) = \alpha + \beta_0 I(T = 0) + \beta_1 I(T = 1) I(X_{j*} > c^*) + \sum_{j=1}^{p} \gamma_j X_j. \tag{2}$$

When the first split is intended to be made at node $T = 0$ instead, the variance accounted for by reference model (Equation 1) has to be compared with the variance accounted for by the following model:

$$E(Y|\mathbf{X}) = \alpha + \beta_0 I(T = 1) + \beta_1 I(T = 0) I(X_{j**} > c^{**}) + \sum_{j=1}^{p} \gamma_j X_j. \tag{3}$$

If a split has been performed, the model associated with this split becomes the new reference model, and the splitting procedures are repeated until no further splits are found or a predefined maximum of splits has been reached. The large regression trunk obtained by STIMA is then pruned using the pruning procedure proposed for Classification and Regression trees.[27]

2.4 | Subgroup identification based on differential effect search

Lipkovich et al[17] proposed a direct search of regions, which have an improved treatment effect. Unlike MOB or IT, SIDES searches only within specific regions of the covariate space and does therefore not estimate the treatment effect for every point in the multidimensional covariate space like the other tree-based methods.

As the other recursive partitioning methods, SIDES starts with the entire data as parent node. But the following procedure differs strongly from MOB, IT, ARDP, and STIMA. SIDES considers all possible splits of the parent node into two child nodes. These pairs of child subgroups have to be ordered in terms of a splitting criterion. Three possible splitting criteria p_1, p_2, p_3 considering the primary efficacy variable were proposed:

1. Maximizing the differential effect between the two child subgroups:

$$p_1 = 2\left[1 - \Phi\left(\frac{|Z_{\text{left}} - Z_{\text{right}}|}{\sqrt{2}}\right)\right]$$

2. Maximizing treatment effect in at least one of the two child subgroups $p_2 = 2\min(1 - \Phi(Z_{\text{left}}), 1 - \Phi(Z_{\text{right}}))$
3. Combination of the first and second criterion $p_3 = \max(p_1, p_2)$.

Z_{left} and Z_{right} denote the test statistics for a one-sided test of the hypothesis of no differential treatment effect in the left and right child subgroups. The cumulative distribution function of the standard normal distribution is denoted by Φ. Smaller values for the splitting criteria indicatie a stronger differential effect between the resulting child nodes. A fourth criterion is based on efficacy and on safety and yields at maximizing the differential effect in terms of those two components.

A pre-specified number M of best child pairs regarding the splitting criterion is selected. Moreover, we retain just the child node exhibiting the larger treatment effect from each pair. These retained subgroups are added to a set of *promising subgroups*. This procedure can lead to child nodes having the same splitting variable but different split points.

After evaluating the splitting criterion for each child node, two further criteria are evaluated on promising subgroups, the selection criterion, and the continuation criterion.

The continuation criterion decides whether a promising subgroup is further split and therefore added to the set of parent nodes. A child node becomes only a part of the set of parent nodes if there is a meaningful improvement of its treatment effect p_C-value compared with the parent treatment effect p_P-value: $p_C \leq \kappa p_P$, where $0 < \kappa \leq 1$ is the relative improvement parameter. Small values κ indicate a very selective identification procedure whereas values close to 1 indicate the opposite. The nodes included in the set of parent nodes are used for further splitting, but SIDES does not consider covariates already used for defining a parent node as potential splitting variables.

The selection criterion has the purpose of deciding whether the promising subgroup can be added to the set of *candidate subgroups*. Subgroups of this set have desirable efficacy. A candidate subgroup is found when the treatment effect P value of a promising subgroup is significant at a one-sided nominal level. This nominal level ϑ can be determined using a permutation-based strategy in order to control the familywise Type I error rate in a weak sense.[17] A Type I error occurs when a promising subgroup is selected as *candidate subgroup*, although, in truth, there is no treatment benefit in any subgroup. For the permutation strategy, the vector of biomarkers $\mathbf{x_i} = (x_{1i}, \dots, x_{ip})$ is randomly permuted against (y_i, t_i) in order to mimic the null data. The SIDES algorithm is then applied to these permuted datasets on a grid of values for ϑ. The proportion of times for which the selection criterion was met for at least one subgroup on the permuted samples is calculated for each of the values for ϑ. The largest value of ϑ for which this proportion does not exceed a pre-specified nominal level is chosen as the significance level for the selection criterion.

The procedure is repeated recursively until none of the identified subgroups meet the continuation criterion, a pre-specified minimum node size, or a pre-specified maximum number of covariates defining the subgroups is met.

2.5 | Adaptive refinement by directed peeling

The ARDP algorithm was originally introduced by LeBlanc et al,[30] and it aims at identifying subgroups of participants with poor prognosis. Patel et al[19] adapted the peeling algorithm for subgroup identification purposes. ARDP peels off fractions of the data in every iteration step. Thus, it is not a recursive-partitioning method. Nevertheless, we can illustrate the algorithm with a tree.

A major difference to the tree-based methods IT, MOB, STIMA, and SIDES is the definition of the cut-off value. Most of the introduced methods perform an exhaustive search for the split. Thus, the sizes of the resulting child node depend on the identified cut-off. In ARDP, it is the other way around. The cut-off value depends on the size of the resulting child node, which we pre-specify by fixing the number of observations to be peeled off in an iterations step. The selection of the splitting variable and its cut-off values are based on several steps. First, we fit the following linear model:

$$E(Y|\mathbf{X}) = \alpha + \beta_0 T + \sum_{j=1}^{p} \beta_j X_i T + \gamma_j X_j.$$

The signs of the interaction effects are used to decide in which direction we peel off the observations. A positive sign for the interaction β_j, $(j = 1, \dots, p)$ means that larger values of the covariate X_j lead to larger treatment effects, therefore, we peel off smaller values in order to increase the treatment effect in the resulting subgroup. A negative sign of β_j leads to peeling off larger values of the corresponding covariate.

We order the observations for each covariate X_j. This ordered list is denoted by $Xj \ast= X_{(1j)}, \dots, X_{(nj)}$. The following peeling step is comparable with the splitting step of the tree-based methods.

The algorithm starts with a subgroup B^0 including all observations and peels off $\max(\alpha_{ardp} \cdot n, n_{\min})$ observations depending on the direction determined with the linear model. The proportion of data to be removed in one iteration step is denoted with α_{ardp} and $n_{\min}$ denotes the minimum number of observations to be peeled off. Thus, we obtain p possible subgroups (B_p^1) in iteration step i. We select the subgroup B_{j*}^i achieving the largest improvement of the treatment effect compared with the effect of the previous chosen subgroup $B^{(i-1)}$ standardized by change in subgroup size. This is repeated recursively until the remaining region includes no less than r observations.

TABLE 1 Overview of the method's properties

	IT	MOB	STIMA	SIDES	ARDP
Aim					
Identifying subgroups defined by predictive covariates	yes	yes	yes	yes	yes
Identifying subgroups defined by prognostic covariates	no	yes	no	no	no*
Algorithm					
Recursive partitioning	yes	yes	yes	yes	no
Evaluating splitting criterion at every possible cut-off point for every covariate	yes	no	yes	yes	no
Selection of covariate and cut-off value simultaneously	yes	no	yes	yes	yes
Covariates can be involved in multiple splits	yes	yes	yes	no	yes
Post-pruning procedure	yes	no	yes	N/A	N/A
Underlying model structure					
Regression model	yes	yes	yes	no	yes
Adjustment for covariate main effects	yes	no	yes	no	yes
- All covariates as main effects	no	no	yes	no	yes
- Dichotomized covariates main effects	yes	no	no	no	no
Results					
Method results in a tree	yes	yes	yes	no	yes
End nodes are the identified subgroups	yes	yes	no	N/A	no
Additional steps needed for obtaining subgroups	no	no	yes	no	yes
Identified subgroups can be overlapping	no	no	N/A	yes	N/A

Note. Since the algorithms are different for the five methods, not every statement can be interpreted reasonably for all methods. These cases are marked with N/A. * The ARDP algorithm initially proposed by LeBlanc et al[30] was developed for identifying prognostic markers only. The extension by Patel et al,[19] which we are using throughout this paper, aims at peeling on predictive markers only.

Abbreviations: ARDP, adaptive refinement by directed peeling algorithm; IT, interaction trees; MOB, model-based recursive partitioning; SIDES, subgroup identification based on differential effect search; STIMA, simultaneous threshold interaction modeling algorithm.

The ARDP algorithm does neither include a pruning nor a selection procedure. But it produces a sequence of nested subgroups of patients benefiting from the experimental treatment. In order to choose one of those subgroups, we need a selection criterion. This is described in Section 3.

2.6 | Discussion on the methods

All five methods try to split the data into subgroups by recursively applying a splitting criterion. This splitting criterion differs across the methods. Furthermore, the tree growing procedure is not identical for all methods considered. IT and MOB are quite similar in their tree growing procedure. Both associate single nodes of the resulting tree with a linear model used for evaluating a splitting criterion. In addition to the splitting criterion, IT and MOB differ in their pruning procedure. IT uses a postpruning procedure, whereas MOB uses pre-pruning in order to prevent nonsignificant branches. STIMA, however, differs not just in the splitting criterion. In this method, the whole tree and not just a node is associated with a linear model. This results in a different interpretation of the end nodes compared with IT and MOB (see Section 3). SIDES in contrast creates multiple trees. For each of the nodes, SIDES decides based on a selection criterion whether the node could be a *candidate subgroup*. Such a *candidate subgroup* is a subgroup, which is likely to have a high benefit from the experimental treatment. SIDES uses a pre-pruning procedure like MOB. But in contrast to MOB, which uses the splitting criterion for both splitting and controlling, the size of nonsignificant branches, SIDES introduces another criterion called continuation criterion.

The peeling procedure of ARDP can be illustrated by a tree with binary splits. But obtaining those splits is different to the recursive partitioning methods. Similar to SIDES, ARDP deletes one of the resulting two child nodes. This results in a tree with just one branch. Since ARDP does not include any pruning or selection step, we cannot interpret the end node as the identified subgroup. Following Doove et al,[21] we give an overview regarding different properties of the methods summarized in a table (see Table 1).

3 | PROPOSAL OF A SUBGROUP CRITERION

In this section, we describe how we obtain subgroups from the results of the five different methods and how we select a potential target population, meaning a subgroup benefiting from the experimental treatment, the *BM+* subgroup, and its complement, the *BM*− subgroup. For this selection, we propose a subgroup criterion. By applying MOB or IT, we get the definition of the identified subgroups directly by the splits of the estimated tree. Figure 1A gives an example for a tree grown by MOB or IT. Four subgroups denoted by S_1, S_2, S_3, and S_4 were identified in this hypothetical example. The subgroup S_1, for example, includes only patients with pretreatment characteristic $X_1 \leq 0$ and $X_2 \leq -0.5$.

Getting the definition of the subgroups by applying STIMA is not as straightforward as it is by using MOB or IT. In order to use STIMA for subgroup identification purposes, we have to force the first split of the tree at the treatment variable. Thus, end nodes only include patients assigned to either the control or the experimental treatment group. We can deduce the subgroup definition from the estimated tree by combining each split defining an end node of the experimental treatment branch with each split defining an end node of the control treatment branch.

The example given in Figure 1B results in three interaction terms included in the linear model: $I(T=0)I(X_1 > -0.5)$, $I(T=1)I(X_2 \leq 0.5)$, and $I(T=1)I(X_2 > 0.5)$. Region R_1 is used as reference. In this example, the resulting subgroups are defined by the following assignment rules:

- S_1: $X_1 \leq -0.5$ and $X_2 \leq 0.5$
- S_2: $X_1 \leq -0.5$ and $X_2 > 0.5$
- S_3: $X_1 > -0.5$ and $X_2 \leq 0.5$
- S_4: $X_1 > -0.5$ and $X_2 > 0.5$

Although this example results in four disjunct subgroups, there is the possibility of obtaining overlapping subgroups by applying this procedure. Splits on the same covariate in the control and experimental subtree cause this. Figure 1D gives an example of a tree produced by STIMA yielding overlapping subgroups. The combinations of R_1 with R_3 and R_1 with R_4 lead to nested subgroups with S_1 defined by $X_1 \leq -0.5$ and $X_2 \leq 0.5$ and S_2 defined by $X_1 \leq -0.5$ and ($X_2 > 0.5$ and $X_1 \leq 0.5$). Another problem arising with this merging procedure is the possibility of obtaining very small subgroups. STIMA does not include any criterion which can avoid this.

Nodes, which are labeled as a *promising* subgroup by the SIDES procedure, are the subgroups we are interested in. The interpretation of those nodes is equivalent to the nodes in MOB and IT. Since SIDES grows multiple trees in each iteration step, it is possible that the identified subgroups are not disjunct.

With the ARDP procedure, we do not get "final" subgroups as it is the case with the other methods. We have to choose a subgroup and its complement from the sequence of subgroups resulting from this procedure. Figure 1C shows a theoretical example of ARDP. We get a sequence of potential *BM*+ subgroups S_1^*, S_2^* and S_3^*. The corresponding *BM*− groups are denoted with *R*.

For defining a future target population, we need a binary classification of patients into a subgroup with an enhanced treatment effect and its complement. Since most of the used methods identify multiple subgroups, we need to dichotomize the results of the subgroup identification methods.

The identified *BM*+ subgroup should be the largest population consisting of identified subgroups exceeding a pre-specified treatment effect threshold denoted by mintrt. Therefore, we calculate the treatment difference $z(\hat{S}) = \mu_1(\hat{S}) - \mu_0(\hat{S})$ in each of the identified subgroups $\hat{S}$. All subgroups identified by the different methods meeting the criteria $\hat{z}(\hat{S}_i) >$ mintrt are amalgamated resulting in the *BM*+ group. The remaining subgroups are merged, and they form the *BM*− group. Note that the procedures use the unadjusted estimate $\hat{z}(\hat{S}) = \frac{1}{|\hat{S}_1|}\sum_{i:x_i\in\hat{S}_1} y_i - \frac{1}{|\hat{S}_0|}\sum_{i:x_i\in\hat{S}_1} y_i$ for the definition of the selected $\widehat{BM+}$ subgroup, where $|\hat{S}_1|$ and $|\hat{S}_0|$ denote the number of treated and untreated patients in an identified subgroup $\hat{S}$. A similar procedure for selecting a target subpopulation was used in Zhao et al[31] and proposed in Lipkovich et al.[14] For ARDP, which produces a sequence of possible $\widehat{BM+}$ subgroups, we choose the largest subgroup meeting the defined subgroup criterion. The $\widehat{BM-}$ is the complement of this chosen subgroup.

The shape of the obtained subgroups can differ across the used methods. In general, the *BM*+ definitions of MOB, IT, STIMA, and SIDES can consist of unions and intersections of different subsets. This can lead to a disjointed definition of the *BM*+ subgroup. Figure 2 illustrates shapes of subgroups of an artifical example. Although there are several tuning parameters or even pruning procedures to keep the tree structure simple, shapes for the *BM*+ population as in subgroup *a* and *c* can be obtained by all considered methods except ARDP. These two shapes might be too complex for application purposes. Subgroup *b* in Figure 2 is jointed and can be obtained by all methods, although ARDP is constructed in such a way, that the generated subgroup is always jointed. The interpretation of a jointed subgroup is easier compared with the other potential shapes.

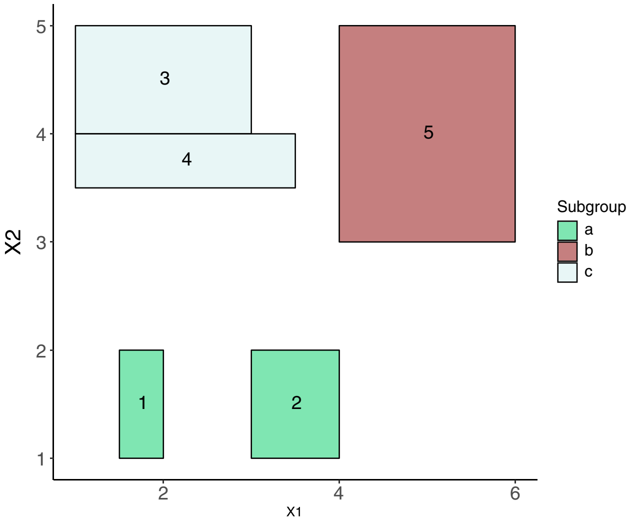


FIGURE 2 Hypothetical example of the shape of subgroups defined by maximal two biomarkers *X*1 and *X*2. Three different shapes are illustrated, which can be obtained by the considered methods. Subgroup a, which consists of the union of the regions 1 and 2 colored in green, can be obtained by all methods except adaptive refinement by directed peeling algorithm (ARDP). Regions 3 and 4 form subgroup b. This shape can also be obtained by all considered methods except ARDP. Subgroup c, obtainable by all five considered methods, consists of just one region, namely, region 5. With regard to application purposes, the shape of subgroup c seems preferable because it just has to be checked whether the measurements of a subjects lie within one box

4 | APPLICATION TO ALS DATA

ALS is a disease affecting the nervous system. Neurons in the brain and the spinal cord controlling the voluntary movement degenerate causing loss of muscle function and paralysis. Therefore, patients with ALS usually have difficulties with chewing, speaking, and swallowing. Since the paralysis can also affect the respiratory system, patients usually die within 3 to 5 years due to respiratory failure.[32]

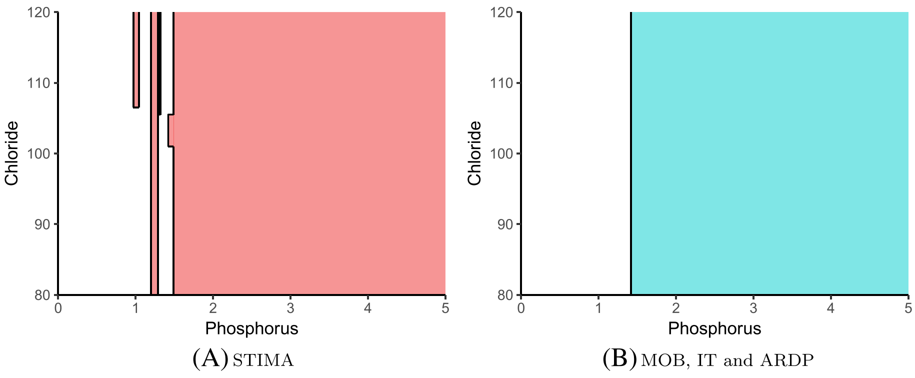
Two treatments for patients with ALS were authorized by the Food and Drug Administration and the EMA, namely, riluzole and edaravone. Edaravone was authorized just recently. Both treatments do not achieve substantial benefit for ALS patients in the overall population.[33,34] Several other compounds for ALS have been developed and investigated as well, but they did not show to be effective.[32] Since there is apparently no treatment with a substantial benefit, it is of interest to investigate whether there is a subgroup of patients with a higher treatment effect. Thus, we apply the five different subgroup identification methods to address this question. Since ALS is designated as orphan disease, clinical trials are relatively small. We used data obtained from the PRO-ACT (Pooled Resource Open Access ALS Clinical Trials) database[35,36] for our analysis. The PRO-ACT database aggregates data from 23 phase II/III trials in order to overcome problems arising during the analysis of clinical data from orphan diseases. The database includes survival times, the ALS functional rating scale (ALSFRS) and forced vital capacity as outcome variables. Baseline variables regarding demographics, laboratory data, vital signs, and family history are available as well. Furthermore, information whether a patient received medication or placebo is available. However, the active treatment is not specified in the data as consequence of de-identification.

We use the ALS data to illustrate the five subgroup identification methods and our proposed criterion in order to identify a *BM*+ subgroup. The ALSFRS is a commonly used measure to evaluate the symptom severity of patients with ALS. It is a score calculated by the sum of 10 assessments regarding the motor function, more precisely speech, salivation, swallowing, handwriting, cutting food and other things, dressing and hygiene, turning in bed, walking, climbing stairs and respiratory. Each of these 10 items referring to different motor functions is rated on a scale from 0 to 4, with 4 indicating a normal function whereas 0 indicates no function. Therefore, the ALSFRS ranges from 0 to 40. ALSFRS-R is a modified version of the ALSFRS containing 12 instead of 10 items. The item regarding the respiratory function in ALSFRS was divided in to three, the others remain the same. Clinical trials for the drug edaravone considered ALSFRS-R after 24 weeks as primary endpoint. The sample size calculation for trial MCI186-19 assumed a difference of 3.0 between the placebo and the edaravone group. The observed difference in trial MCI186-19 was 2.49 with a 95% confidence interval (CI) of 0.99-3.98.[33] Since the number of observations of ALSFRS-R after approximately 6 months is smaller than for ALSFRS, we used ALSFRS as outcome variable in our analyses. Since a difference of 3.0 score points on the ALSFRS-R is relatively a smaller difference as 3.0 points on the ALSFRS scale, we think that 3 is a reasonable threshold for our proposed subgroup criterion. We use the ALSFRS after approximately 6 months allowing a window of 20 days as outcome variable. In order to keep the presentation simple, we preselected two covariates. This preselection is based on a linear model. We included all available continuous covariates measured at baseline as main effects and as treatment-by-covariate interaction. This includes the covariates ALSFRS at baseline and the forced vital capacity. We have chosen the two variables with the smallest *P* values for their interaction effects, namely, phosphorus and chloride. Chloride is a well known prognostic factor. Lower serum chloride levels are associated with a worse prognosis.[37] Phosphorus, however, is not mentioned as a potential prognostic marker in literature. Moreover, no substantial correlation of these two markers is observable in the ALS data ($\rho = -0.06$; 95% CI, −0.10 to −0.02).

The sample size of the dataset used for the pre-selection and the subgroup identification methods included n=2156 full observations. We have used the same tuning parameters for the subgroup identification methods as for the simulation study. The used tuning parameters are listed in Section 5.1.

SIDES identifies three candidate subgroups. Since the treatment effects in these candidate subgroups are smaller than the prespecified threshold for the subgroup criterion, no BM+ subgroup is identified by this procedure. IT and STIMA do not identify a *BM*+ subgroup when the subgroup criterion is applied after the originally proposed pruning procedure. However, we have to differentiate those two results from one another, eg, IT's initial tree is pruned back to the root using the originally proposed pruning procedure, whereas STIMA's pruned tree involves some splits. Since the estimated treatment effects in these resulting subgroups do not exceed the specified threshold, no target population is identified. Despite this difference, we apply the subgroup criterion on both initial trees. Then, all methods, besides SIDES, identify a *BM*+ subgroup. The *BM*+ subgroup identified by MOB, IT, and ARDP is defined by the variable phosphorus, and their cut-off values differ only slightly. MOB's and IT's *BM*+ subgroup only includes patients with values larger than 1.42 for the phosphorus variable and ARDP uses the value 1.36 as cut-off. Since those identified target subgroups are largely overlapping, we illustrated the shape of IT's, MOB's and ARDP's target subgroup in a single plot, see Figure 3. STIMA identifies a *BM*+ group consisting of the union of five subsets. One of these subsets includes patients with values greater than 1.49 for the covariate phosphorus. This subset overlaps with the other *BM*+ subgroups identified by the other methods. Moreover, STIMA's *BM*+ group involves patients with both 1.42< phosphorus ≤ 1.495 and 100.95<chloride≤ 105.5 and patients with both 1.30<phosphorus ≤ 1.32 and chloride≤ 105.5. Furthermore, patients with 1.20< phosphorus ≤ 1.29 and patients with both 0.97<phosphorus ≤ 1.05 and chloride> 106.5 are included in the target subgroup of STIMA. The subgroup selected by STIMA is the largest one. It includes 483 patients, whereas ARDP involves 204 patients and both MOB and IT 145 patients. Nevertheless, these four methods identify similar benefiting subgroups. Patients with larger values for the covariate phosphorus seem to profit the most. The tress obtained by the different methods are illustrated in Appendix A.

As it was mentioned in the artificial example illustrated in Figure 2, the *BM*+ subgroup does not necessarily consist of jointed subsets. This is also the case for STIMA's identified *BM*+ subgroup. Note that this complex shape of STIMA is a result of using the initial tree with a high number of allowed splits (see Section 5.1). The pruned tree of STIMA involved only a single split at the active treatment arm.



(A) STIMA (B) MOB, IT and ARDP

FIGURE 3 Shape of the selected *BM*+ subgroups for amyotrophic lateral sclerosis (ALS) patients. Since the selected *BM*+ group of adaptive refinenement by directed peeling algorithm (ARDP), interaction trees (IT), and model-based recursive partitioning (MOB) is almost the same, it is represented as a single region. Simultaneous threshold interaction modeling algorithm (STIMA)'s identified BM+ subgroup consists of the union of multiple subgroups defined by chloride and phosphorus or only by phosphorus. The BM+ subgroup of IT, MOB, and ARDP in contrast is just defined by phosphorus and one corresponding cut-off value. Therefore, we obtain a half-open box shape for IT's, MOB's, and ARDP's BM+ subgroup

5 SIMULATION STUDY

In the following section, we investigate the performance of the subgroup identification methods IT, MOB, STIMA, SIDES, and ARDP for the identification of a *BM*+ in connection with our proposed subgroup criterion using the Clinical Scenario Evaluation framework.[24] In Section 5.1, we specify the used tuning parameters for the five methods. Then, we define criteria like the selection accuracy, sensitivity, and specificity, which are used to evaluate the performance of the methods. Afterwards, we introduce the considered data-generating mechanisms and the results.

5.1 Options: Subgroup identification methods and their tuning parameters

We can specify different tuning parameters for the methods used including the subgroup criterion. In the following we describe how we have chosen those parameters in our simulation study.

5.1.1 Interaction trees

The R code for IT is provided by the authors and is available at http://biopharmnet.com/subgroup-analysis-software/.

The R implementation requires a learning and test sample in order to overcome the overoptimism induced by the greedy search for selecting the best sized tree. We use 80% of the dataset as learning sample and the remaining 20% as test sample. Therefore, IT uses less observations for the growing process of the initial tree compared with the other methods. We set the parameter controlling the penalization of additional splits in the pruning procedure to the value $ln(n)$, which is said to provide the best subtree selection compared with other values for this complexity parameter. The minimum terminal node sizes was set to to 20, the minimum number of observations in one of the the treatment arms for all permissible splits is set to 5 and the maximum depth of tree was chosen to be 15.

5.1.2 Model-based recursive partitioning

MOB for subgroup identification is implemented in the package `partykit`.[29,38] We used $E(Y|X, T) = \beta_0 + \beta_1 T$ as model and all available covariates $X_1, \dots, X_p$ as possible splitting variables. The minimum node size was set to 20 and the maximum depth of the tree was set to 15. For the other tuning parameters the default values were used.

5.1.3 Simultaneous threshold interaction modeling algorithm

The R package `stima` provided by the authors[39] was used for applying STIMA. For the minimum size of a terminal node, we used the value 20 and for the maximum number of splits we used 15. The other tuning parameters both for the growing and the pruning procedure were set to their default values.

5.1.4 Subgroup identification based on differential effect search

The R code developed by the authors and published on http://biopharmnet.com/subgroup-analysis-software/ was used for SIDES. The minimum subgroup size was chosen to be 20, and both the maximum number of covariates used for defining the subgroups and the number of retained subgroups for each parent were set to the value 3. The relative improvement parameter was set to 1. The local multiplicity adjustment used for penalizing covariates with a large number of candidate splits implemented in the available R code was chosen to be 0.1. Moreover, we used 500 permutation for computing the significance level for the selection criterion. An identified subgroup was selected as candidate subgroup if both the *P* value of the treatment effect in the identified subgroup and its adjusted treatment effect *P* value based on the resampling method were smaller than 0.05.

5.1.5 Adaptive refinement by directed peeling algorithm

For the ARDP algorithm, we used a self-implemented R code. The tuning parameters α and r were set to the values 0.1 and 20, respectively. The minimum terminal node size was set to 20.

5.1.6 Subgroup criterion

The threshold for the subgroup criterion is chosen based on the assumptions for generating the datasets.Therefore, the chosen values are presented in Section 5.2.

5.2 | Assumptions: Models and parameter values assumed in the simulations

In this section, we present the data generating models. Each dataset consists of a continuous response Y, a binary treatment variable T and p covariates for each subject $i = 1, \dots, n$. The treatment variable T is drawn from a binomial distribution $\mathcal{B}(1, 0.5)$ and the covariates $\mathbf{X} = (X_1, \dots, X_p)$ from $\mathcal{N}_p(0, \sigma^2)$. For simplicity, we chose $\sigma^2 = I_p$, with I_p denoting the identity matrix. The outcome Y is generated by using $Y_i = \mu(T_i, \mathbf{X}_i) + \epsilon_i$ with $\epsilon_i \sim \mathcal{N}(0, 1)$. As in Graf et al,[40] we considered a step function model (M1) and a linear trend model (M2) as mean functions $\mu(T, \mathbf{X})$, ie,

$$\mu(T, \mathbf{X}) = 0.2 \cdot T + \gamma \cdot I(X_1 > c) + \beta_1 \cdot T \cdot I(X_1 > c), \tag{M1}$$

$$\mu(T, \mathbf{X}) = 0.2 \cdot T + \gamma \cdot X_1 + \beta_1 \cdot T \cdot X_1. \tag{M2}$$

We vary the parameters as follows:

- Number of covariates p: $p = 1$ and $p = 4$; The performance of all methods in settings with $p = 1$ is slightly better compared with those with three noise covariates. For ARDP, we can observe a stronger influence of the noise covariates on the classification of patients. Since the setting without noise covariates is less relevant in praxis, we do not discuss the results.
- Cut-off value c in model M1: $c = 0$, $c = -0.5$ and $c = 0.5$. The choice of the cut-off value influences the size of the true $BM+$ subgroup. For $c = 0$, the true subgroup involves around $n/2$ subjects, when the threshold mintrt is chosen appropriately.
- Sample size: $n = 600$, $n = 1200$, and $n = 2400$
- Prognostic effect size: $\gamma = 0$, $\gamma = 0.2$, and $\gamma = -0.2$
- Predictive effect size referring to the difference in outcome between the experimental and control treatment: We considered a small ($\beta_1 = 0.3$), medium ($\beta_1 = 0.5$), and large ($\beta_1 = 1$) effect. In order to evaluate the Type I error rate, we also considered cases with no predictive effect ($\beta_1 = 0$).

Furthermore, we adapted one of the mean functions used in Dusseldorp et al[18] and the mean function presented in Lipkovich et al,[17] namely,

$$\mu(T, \mathbf{X}) = \gamma_1 \cdot X_1 + \dots + \gamma_p \cdot X_p + \beta_1 \cdot T \cdot I(X_1 > 0), \tag{M3}$$

$$\mu(T, \mathbf{X}) = a \cdot \left\{ I(X_1 > 0) \cdot (1 - \frac{n_{01}}{n}) - I(X_1 \leq 0) \cdot \frac{n_{01}}{n} \right\} T. \tag{M4}$$

Model M3 includes the main effects of all splitting covariates. The number of patients having values larger than 0 for covariate X_1 is denoted by n_{01} in model M4. This mean function adapted from Lipkovich et al includes in contrast to the other mean functions a qualitative interaction. Moreover, the overall mean is 0 in this setting. For models M3 and M4, we used the following parameter values:

- Number of covariates $p = 4$
- Sample size: $n = 600$, $n = 1200$, and $n = 2400$
- Prognostic effect sizes: $\gamma_1 = -0.3$, $\gamma_2 = 0.4$, $\gamma_3 = 0$, and $\gamma_4 = 0.3$
- Predictive effect size referring to the difference in outcome between the experimental and control treatment: A medium ($\beta_1 = 0.5$) and large ($\beta_1 = 1$) effect were considered for model M3 and $a \in \{1, 2\}$ for model M4.

For each parameter combination, we generate 500 datasets. The threshold for the subgroup criterion is set to the value mintrt $= 0.4$ for all considered parameter combinations. Since the treatment effect increases with increasing values for the biomarker X_1 in settings with a linear trend (model M2), the $BM+$ subgroup depends strongly on the mintrt value. Therefore, we varied the threshold mintrt $= 0.4., 0.6, \dots, 1.2$ only for the linear trend setting.

5.3 | Metrics: Operation characteristics

The subgroup criterion has the same purpose as companion diagnostics. Both distinguish patients who will respond better to experimental treatment from those who will respond less. Therefore, we want to describe their analytical accuracy with the following criteria[26]:

- Sensitivity: P(patients are assigned to the $\widehat{BM+}$ group | patients truly belong to the $BM+$ group)

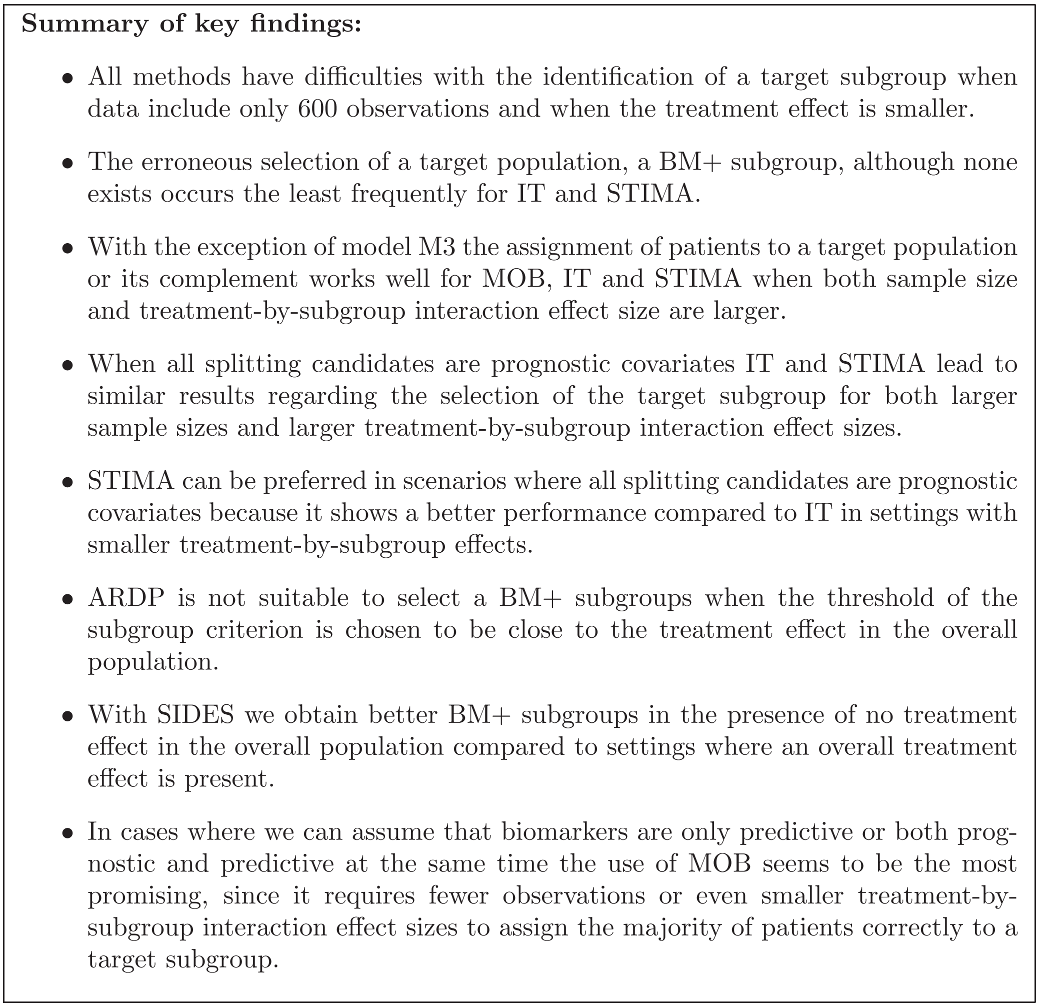
- Specificity: P(patients are assigned to the $\widehat{BM-}$ group | patients truly belong to the $BM-$ group)
- Selection Accuracy: P(patients are correctly classified)
- Type I error rate
- Type II error rate

A Type I error occurs when an estimated $\widehat{BM+}$ group, which does not include all or no patients, is identified although no subgroups are present meaning the treatment effect is homogeneous across the entire population. This corresponds to falsely rejecting the null hypothesis that no subgroup is present. In contrast, the Type II error is defined as wrongly retaining the null hypothesis. Therefore, the Type II error refers to the situation when the $\widehat{BM+}$ group includes all or no patients although “real" subgroups are present. Note that we actually do not test any hypotheses, but we will still refer to the situations described above as Type I and Type II error. For the calculations of the Type I error rates, we consider the step function model without any predictive effect. We calculated the proportions of datasets for which a target subgroup (not equal to the overall study population or the null set) was found.

5.4 | Results

The simulation study compared the performance of the five subgroup identification methods in terms of their ability to identify BM+ and BM− subgroups as defined in Section 3. We summarize the findings of our investigations in Figure 4.

The Type I error rates of the five subgroup identification methods are shown in Figure 5. SIDES' Type I error rate decreases with increasing sample size and seems not strongly influenced by the presence of a prognostic effect. For MOB,



Summary of key findings:

- All methods have difficulties with the identification of a target subgroup when data include only 600 observations and when the treatment effect is smaller.
- The erroneous selection of a target population, a BM+ subgroup, although none exists occurs the least frequently for IT and STIMA.
- With the exception of model M3 the assignment of patients to a target population or its complement works well for MOB, IT and STIMA when both sample size and treatment-by-subgroup interaction effect size are larger.
- When all splitting candidates are prognostic covariates IT and STIMA lead to similar results regarding the selection of the target subgroup for both larger sample sizes and larger treatment-by-subgroup interaction effect sizes.
- STIMA can be preferred in scenarios where all splitting candidates are prognostic covariates because it shows a better performance compared to IT in settings with smaller treatment-by-subgroup effects.
- ARDP is not suitable to select a BM+ subgroups when the threshold of the subgroup criterion is chosen to be close to the treatment effect in the overall population.
- With SIDES we obtain better BM+ subgroups in the presence of no treatment effect in the overall population compared to settings where an overall treatment effect is present.
- In cases where we can assume that biomarkers are only predictive or both prognostic and predictive at the same time the use of MOB seems to be the most promising, since it requires fewer observations or even smaller treatment-by-subgroup interaction effect sizes to assign the majority of patients correctly to a target subgroup.

FIGURE 4 Key simulation findings for the performance of five subgroup identification methods in terms of their ability to identify BM+ and BM− subgroups

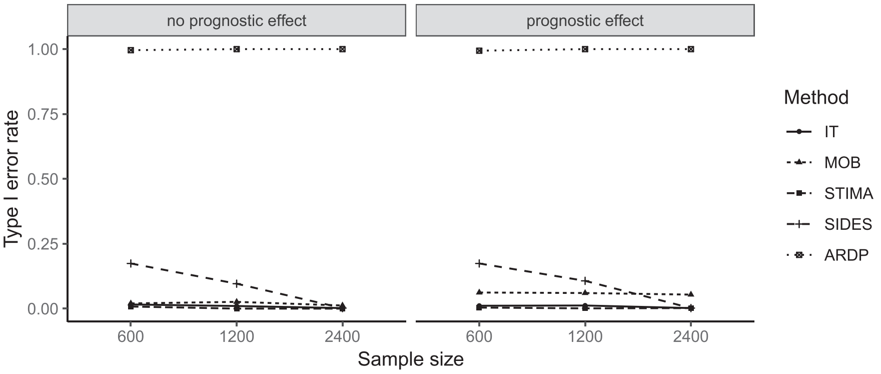


FIGURE 5 Type I error rate for data generated with model M1 as mean function, $\beta_1 = 0$ and $c = 0$. The y-axes represent the Type I error rates, whereas the x-axes represent the considered sample sizes. The left panel includes the Type I error rates when the prognostic effect γ is set to $\gamma = 0$. The direction of the prognostic effect does not influence the Type I error rate. Therefore, the right panel corresponds to the Type I error rates of the settings with $\gamma = 0.2$ or $\gamma = -0.2$

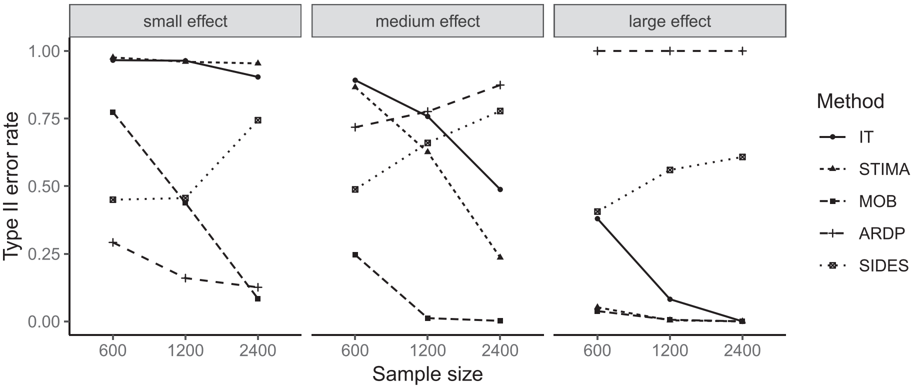


FIGURE 6 Type II error rate for settings with the step function model (M1), $\gamma = 0$ and $c = 0$. For simultaneous threshold interaction modeling algorithm (STIMA) and model-based recursive partitioning (MOB) low Type II error rates are observable when the predictive effect is large. Interaction trees (IT)'s error rate depends strongly on the sample size. For medium effect sizes, MOB clearly outperforms the other methods in terms of the Type II error rate. Adaptive refinement by directed peeling algorithm (ARDP) has the lowest Type II error rate in settings with small effects and small sample sizes, which is due to the treatment effect in the overall population and the chosen mintrt value. MOB's Type II error rate decreases strongly with the increase of the sample size when the effect size is small

STIMA, and IT, we can observe small Type I error rates. Since MOB does not just look for instabilities in the partial score function of the treatment parameter but also of the intercept, it is not surprising that the Type I error rates of MOB are affected by the presence of prognostic effects. When prognostic effects are present, MOB is more likely to identify a subgroup even though none is present. STIMA's and IT's Type I error rates differ only slightly across the settings with different prognostic effect sizes and different sample sizes. Both methods achieve Type I error rates below 2 %. Nevertheless, STIMA is the method with the smallest Type I error rate. Figure 5 shows that ARDP does not control the Type I error. ARDP always results in the selection of a target subgroup with the exception of two situations: (a) when the estimated treatment effect in the overall population is larger than the threshold for the subgroup criterion and (b) when no subgroup of the obtained sequence exceeds the threshold. These situations do not depend on the presence of true subgroups.

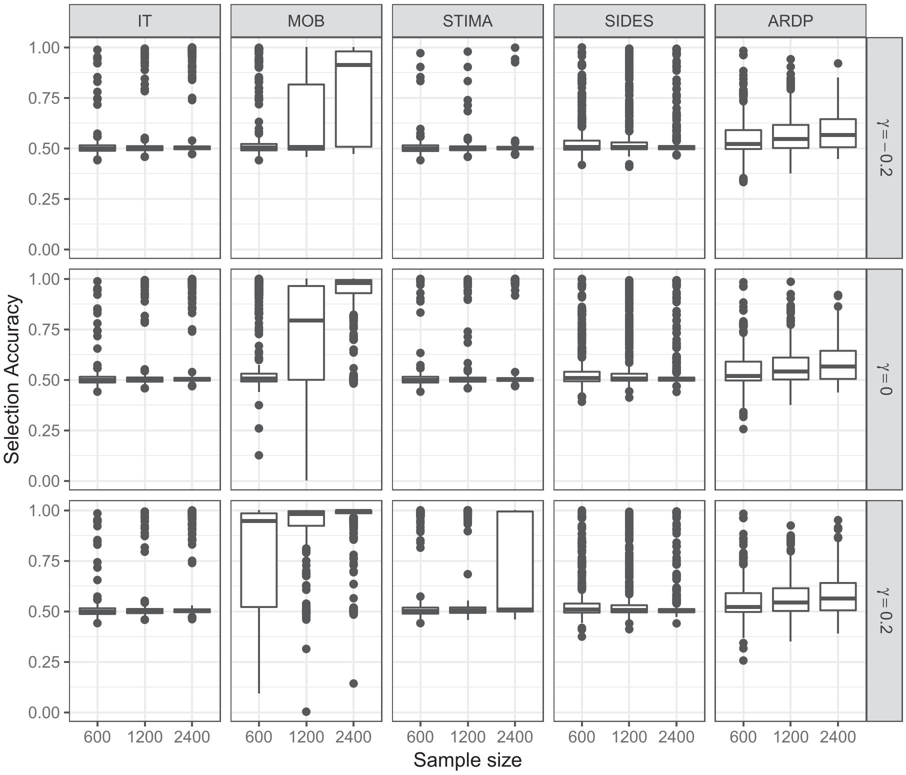
The Type II error rate for settings with model M1 in which no prognostic effect is present and where the *BM*+ and *BM*− subgroups are of equal size is presented in Figure 6. For large interaction effects in model M1, we can observe low Type II error rates for MOB and STIMA. These methods do not just have low Type II error rates when the interaction effect is large but they do also classify nearly all patients correctly. In the presence of a large predictive effect, the influence of prognostic effects, the size of the true *BM*+ subgroup, and the sample sizes were negligible for MOB and STIMA in terms of the selection accuracy. IT achieves a comparable performance in settings with large interaction effects when the dataset is large enough. For IT, we can observe that the sample size is important for its Type II error performance. Recall that the total sample is split into 80% training sample and 20% test sample for IT. Both sizes should be large for a good performance of IT, eg, with small training samples it is likely to not identify the correct splits. Important splits identified in the initial tree are prone to be pruned back due to a small test sample.

When the biomarker is not only predictive but also prognostic ($\gamma \neq 0$), the performance of MOB and STIMA in small or medium predictive effect settings is affected (see Figure 7). When predictive and prognostic effects have the same direction as it is the case when γ is set to 0.2, the performance improves. For this case and $\beta_1 = 0.3$ and $n = 600$, MOB achieves a median selection accuracy of 0.98. In case of different directions, only patients from the subgroup with an expected lower outcome benefit from the treatment. In this setting, the performance deteriorates, eg, MOB does not identify any subgroup in the scenario with $\beta_1 = 0.3$ and $n = 600$. The reason for MOB's dependency on the presence of a prognostic effect is explained by its splitting criterion, which wants to avoid missing important cut-off by investigating not only the instability of the treatment effect. Seibold et al[16] showed how the choice of a prognostic effect affects the partial score functions and therefore the subgroup identification. In cases where the treatment difference between the $BM+$ and $BM-$ is not large ($\beta_1 = 0.3$), IT and STIMA have difficulties identifying any subgroup. This can be seen in Figure 7, which presents the selection accuracies for settings with model M1 depicted by boxplots. Usually, a selection accuracy of 0.5 refers to a random classification. This is not the case for scenarios with $c = 0$, what can be verified by looking at the sensitivity and specificity or the Type II error. The selection accuracy of 0.5 refers to a selection of the overall population as $\widehat{BM+}$ or $\widehat{BM-}$ subgroup. The reason for this is the size of the true $BM+$ subgroup. Based on the value $c = 0$ and mintrt $= 0.4$, the size of the true underlying $BM+$ subgroup is $n/2$. IT's and STIMA's difficulty of identifying subgroups becomes more evident by looking at the Type II error. The Type II error rate for these two methods is near to one (eg, see Figure 6) when the treatment effect difference between the $BM+$ and $BM-$ group is small. Although, the power for the two-sided heterogeneity test[41] is larger than 95% assuming a sample size $n = 2400$ and the standard two-sided significance level of 0.05.

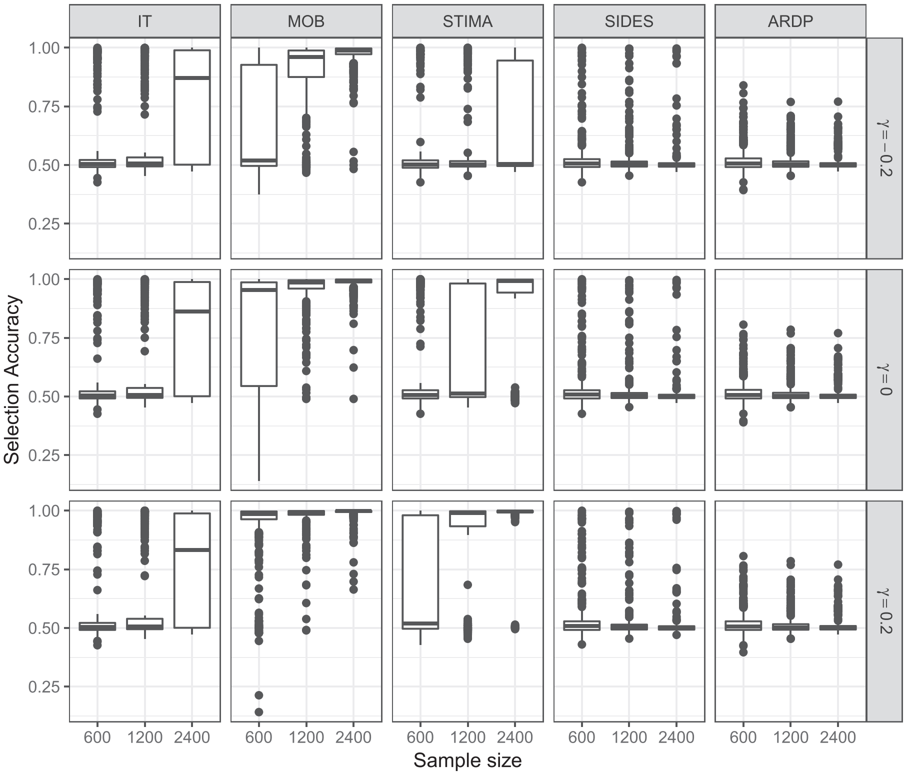
For both methods, a subgroup identification is observable in their initial trees but the applied pruning procedures prune the initial trees back to their root. Since our proposed subgroup criterion can be seen as a pruning rule, we additionally investigated the performance of our subgroup criterion applied on the initial trees of IT and STIMA. This means that the large trees of IT and STIMA are grown as proposed by the developers. But instead of the originally proposed pruning procedures, we apply the subgroup criterion described in Section 3 on the subgroups obtained by the initial trees. The pruning procedures for IT and STIMA aim at reducing the amount of unreliable branches in the decision tree. Therefore, it is not surprising that the Type I error increases when we apply the subgroup criterion to the large initial trees. Using the subgroup criterion on the unpruned STIMA tree results in a Type I error rate of 1, whereas IT's Type I error is larger than 0.6 for all considered parameter combinations. IT's Type I error rate increases with increasing sample size because numerous small subgroups are identified without the originally proposed pruning procedure. Reducing the complexity of the initial tree by setting the maximum tree depth to a smaller value, namely, to 5, did decrease the Type I error rate for STIMA and IT considerably. The Type I error is close to 0 for those cases. Although changing the tuning parameter referring to the maximum tree depth has a huge effect on the Type I error rate, the improvement achieved by reducing the tree complexity on the selection accuracy is moderate compared with allowing more complex trees. Nevertheless, the use of the subgroup criterion on the unpruned trees regardless of the chosen maximum tree depth criterion improves the selection accuracy of IT and STIMA compared with the accuracy of the pruned trees in settings with small and medium interaction effects (see Figure 8). Imposing an even smaller value for the tree depth would improve the performance, since the obtainable complexity of the trees is closer to the complexity of the true underlying tree. For larger effects and sample sizes, we classify more patients correctly when we apply the subgroup criterion after the originally proposed pruning procedures.

With the chosen value for the relative improvement parameter, SIDES identifies large sets of subgroups. When multiple identified subgroups fulfill the subgroup criterion the resulting $\widehat{BM+}$ is very large or can even be equivalent to the overall study population. This behavior can be observed in the results of the simulation study, regardless of the presence of a prognostic effect or the true effect size. We can observe this peculiarity in Figures 6 and 7. The Type II error rate is not equal to 1, and the selection accuracies of SIDES are around the value 0.5 for settings with equal $BM+$ and $BM-$ sizes. This indicates that the size of identified subgroups differs only slightly from the sample size. SIDES tends to identify several subgroups defined by noise covariates only. The true treatment effect in these spurious subgroups is equal to the overall treatment effect, which exceeds the chosen threshold for the subgroup criterion in some settings of the simulation study. Therefore, the subgroup criterion is not able to separate the spurious subgroups from subgroups defined by the predictive biomarker (X_1), what results in too large BM+ subgroups.

ARDP in general chooses the overall study population as $\widehat{BM+}$ group when the treatment effect in the overall population is larger than the pre-specified mintrt threshold. This is the case for the step function model (M1) with $\gamma = 0$, $c = -0.5$,

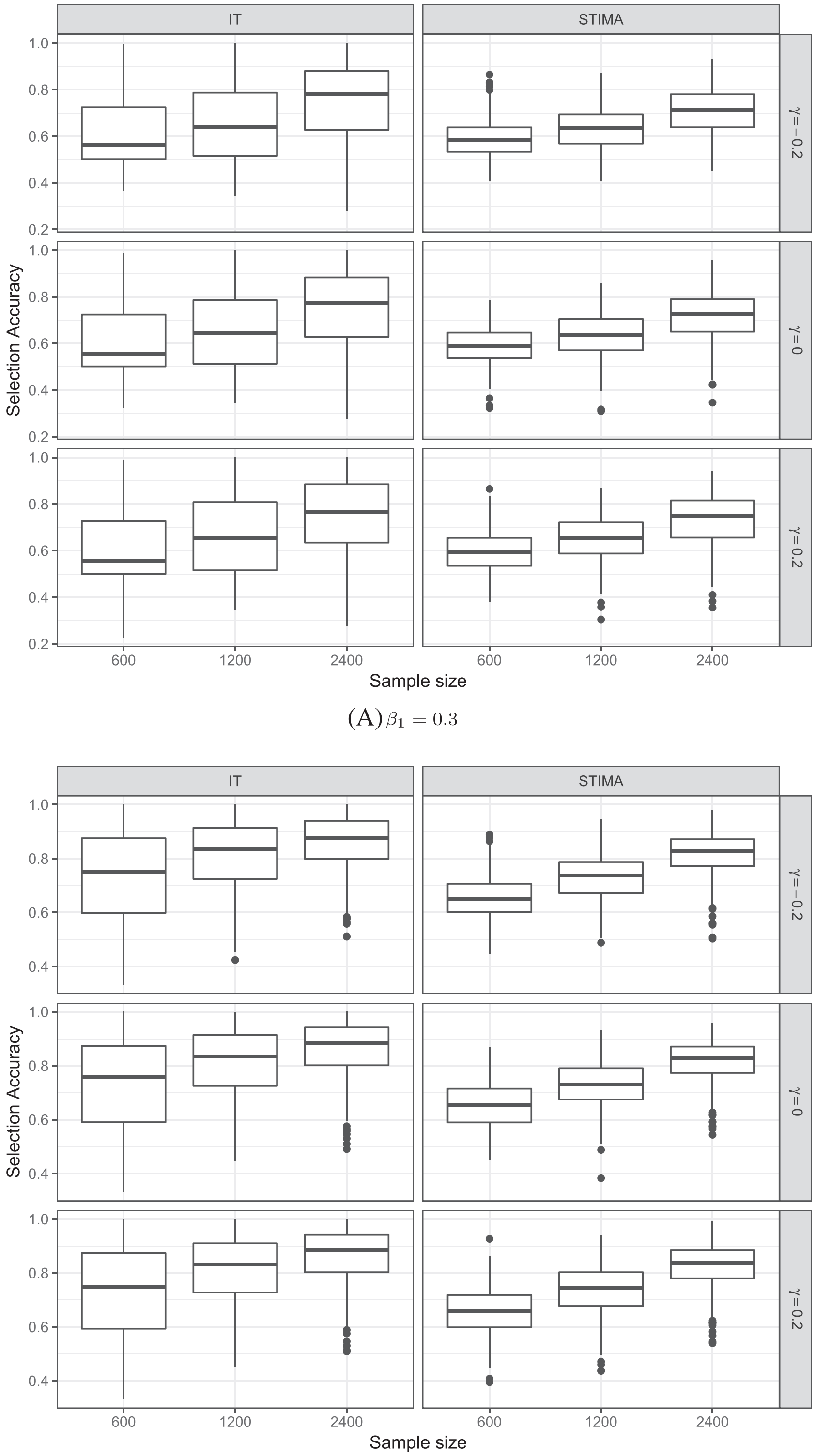


(A) $\beta_1 = 0.3$



(B) $\beta_1 = 0.5$

FIGURE 7 Selection accuracy for data generated by using model M1 with $c = 0$ and A, $\beta_1 = 0.3$ or B, $\beta_1 = 0.5$. The median is displayed with the horizontal line within the box. For several setting and methods, eg, IT for all settings shown in A, median, upper, and lower quartile differ only slightly or are even equal. Therefore, some of the shown boxes are just a horizontal line



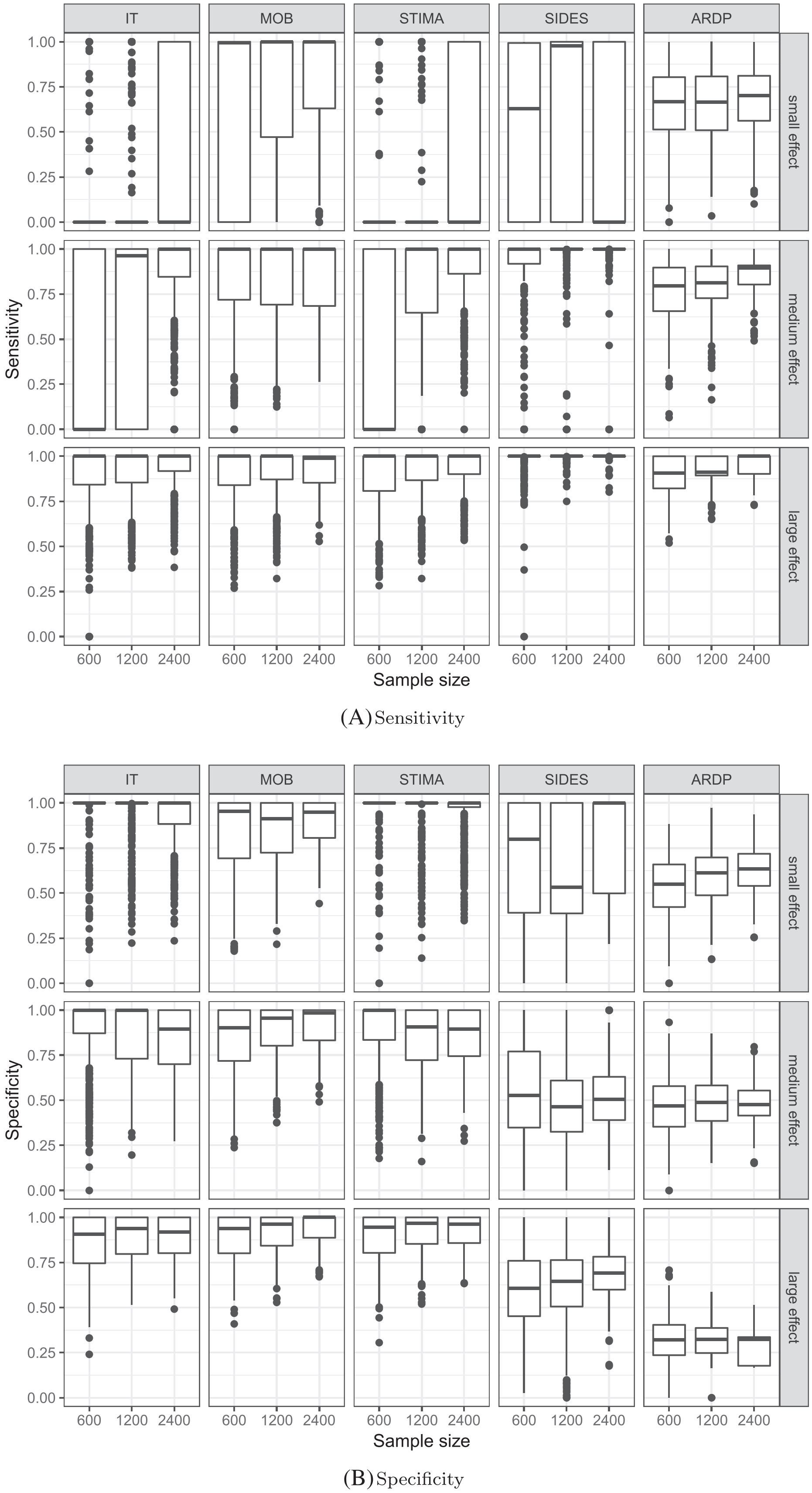
(A) $\beta_1 = 0.3$

(B) $\beta_1 = 0.5$

FIGURE 8 Selection accuracy for interaction trees (IT) and simultaneous threshold interaction modeling algorithm (STIMA) without the originally proposed pruning procedures using a maximum tree depth of 5 in settings with model M1, $c = 0$ and A, $\beta_1 = 0.3$ or B, $\beta_1 = 0.5$

and all considered interaction effects or for $\gamma = 0$, $c = 0$ and a medium $\beta_1 = 0.5$ or large $\beta_1 = 1$ interaction effect. For a small interaction effect, the true treatment effect in the overall population is 0.35. Therefore, some peeling steps are performed, and the identified $\widehat{BM+}$ subgroup does not involve all subjects. Therefore, the specificity is not equal to zero in those cases. But the performance of ARDP is still poor.

In settings with model M1, the size of the true underlying subgroup is the same across the different values for the interaction effect β_1 when the subgroup criterion mintrt is fixed to the value 0.4. For model M2, this is not the case. For



(A) Sensitivity

(B) Specificity

FIGURE 9 Sensitivity and Specificity using model M2 with $\gamma = 0$

a fixed subgroup criterion mintrt, the size of the true $BM+$ decreases with a decreasing interaction effect. Figure 9 shows the sensitivity and specificity for all five methods in settings with a continuous treatment-by-biomarker interaction and no prognostic effect. As in the setting with data generated by the step function model (M1), we can observe that MOB classifies patients better compared with the other methods. In contrast to the scenario with the step function, the influence of a prognostic effect seems to be negligible for all methods. SIDES in general seems to select too large subgroups, and therefore we can observe a high sensitivity but a low specificity. When a continuous treatment-by-biomarker is present, SIDES performance improves with increasing sample sizes. As SIDES, the ARDP algorithm's performance improves with increasing sample size, and it also tends to select too large $\widehat{BM+}$ subgroups. But as can be seen in Figure 9, not all patients belonging to the true $BM+$ subgroup are included in the identified $\widehat{BM+}$ subgroup. This can be explained by ARDP's sensitivity towards the presence of noisy covariates.

Model M3 includes in contrast to the step function (M1) and linear trend (M2) model all available covariates as main effects. Furthermore, the main effects are not based on dichotomized covariates. STIMA, ARDP, and IT adjust for main effects. However, only STIMA and ARDP consider the main effects of all available covariates in their underlying regression models. Moreover, these main effects of underlying regression model are not based on dichotomized covariates.

Figure 10 shows that IT's median selection accuracy is 0.5 for a medium treatment effect and all considered sample sizes. As in settings with data generated with model M1 and $c = 0$, a selection accuracy of 0.5 means that all patients are assigned to either the BM+ or BM− subgroup. For ARDP and MOB, the median selection accuracy slightly increases with an increasing sample size in settings with a medium effect size. However, the median selection accuracies of these two methods are smaller than 0.7. SIDES' median selection accuracy is the highest for a medium treatment effect and 1200 observations, but its variation is larger than the variation of MOB's and ARDP's selection accuracies. For 2400 observations and a medium treatment effect, the highest median selection accuracy is observable for STIMA. In almost all of the performed simulation runs, all patients are classified correctly to either the BM+ or BM− subgroup by STIMA when the treatment effect difference is large ($\beta_1 = 1$). For IT, we can observe almost the same when more than 1200 observations are available, although the variation of the selection accuracies is slightly larger compared with STIMA's variation. MOB's median selection accuracy is always below the ones of IT and STIMA in settings with a high treatment effect. This is due to the presence of multiple prognostic covariates, which MOB does not adjust for. In fact, MOB approximates these prognostic effect by step-functions, which are induced by splitting the tree on these covariates. As observed with data generated with M1, ARDP and SIDES tend to assign all patients to either BM+ or BM− when the treatment effect difference is set to be 1. For both methods, this can be explained by the overall treatment effect and the value for the subgroup criterion as in settings with model M1.

The previous models used for data generation included a quantitative treatment-by-biomarker interaction. Model M4 generates data with a qualitative treatment-by-biomarker interaction and an overall mean of 0. Figure 11 shows that IT, MOB, STIMA, and SIDES selection accuracies are close to 1 in 75% of the simulation runs for data generated with model M4 and $a = 2$. Even for the settings with a smaller treatment difference, namely, those with $a = 1$, the assignment of patients to the BM+ or BM− subgroup is good for IT, MOB, STIMA, and SIDES. The median selection accuracies are close to 1, and we can just observe a larger variation for IT and SIDES in settings with 600 observations. ARDP's selection accuracy is smaller compared with the accuracies obtained by applying the other methods. However, we can observe that ARDP assigns more patients correctly with increasing sample size.

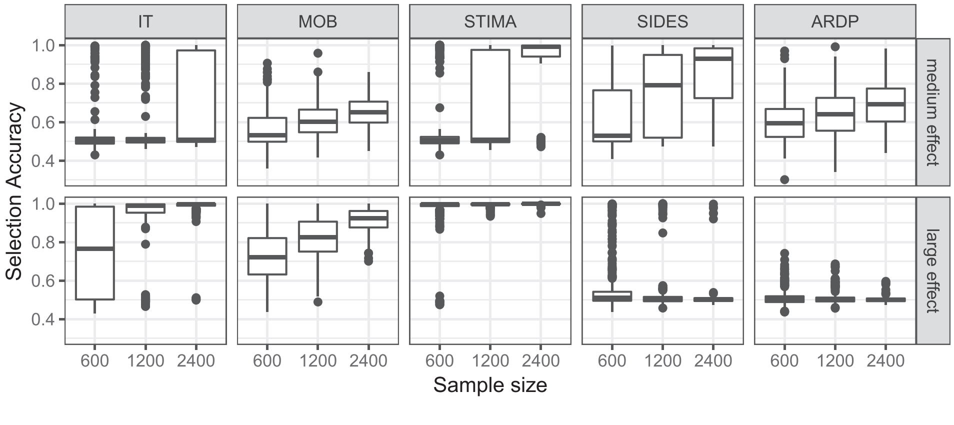


FIGURE 10 Selection accuracy for data generated with model M3. Simultaneous threshold interaction modeling algorithm (STIMA) classifies all patients correctly to either the BM+ or BM− subgroup for settings with a larger interaction effect ($\beta_1 = 1$)

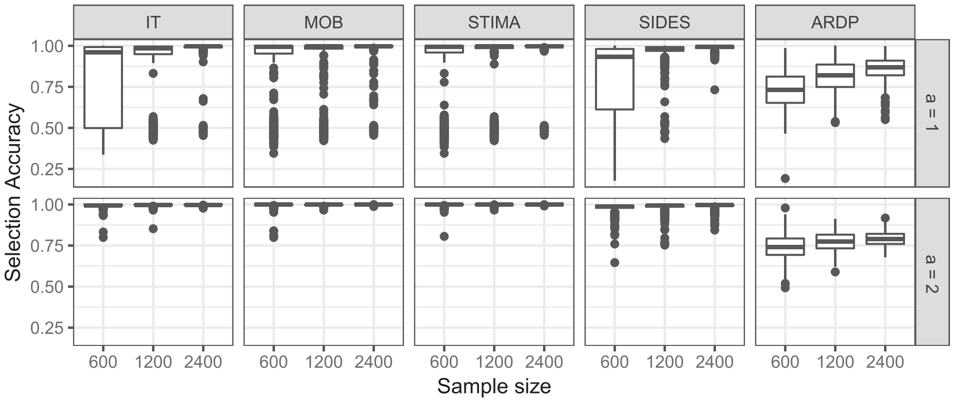
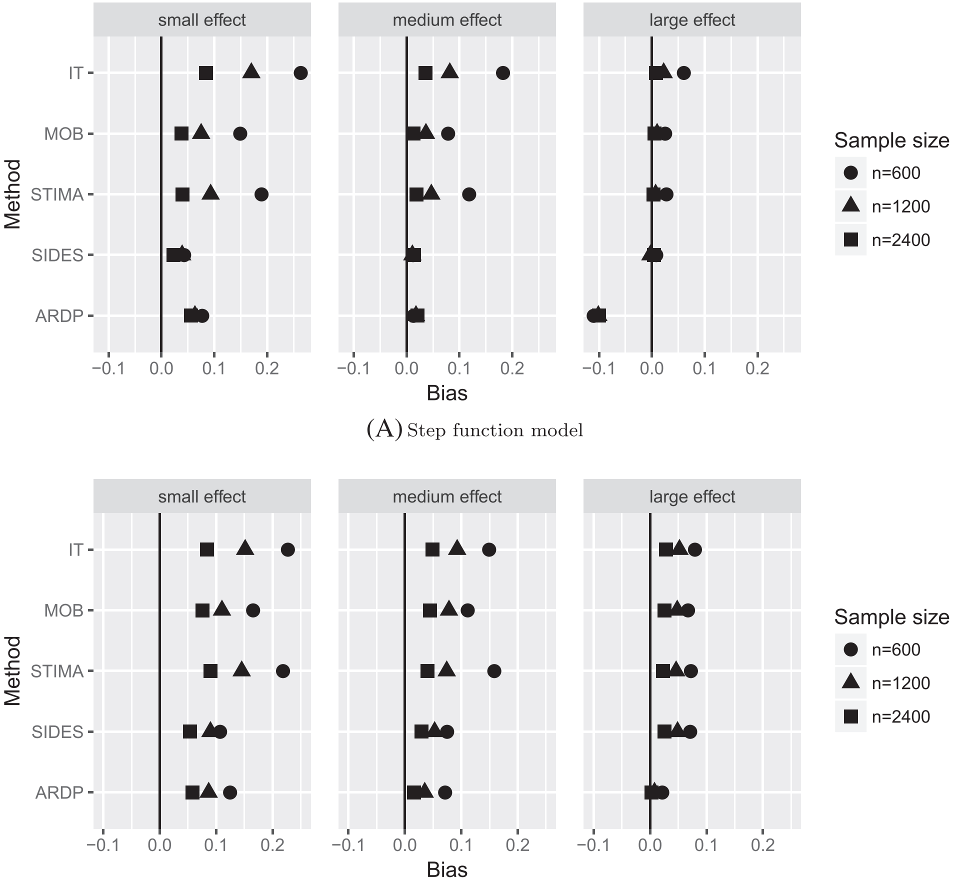


FIGURE 11 Selection accuracy for data generated with model M4 including a qualitative interaction



(A) Step function model

(B) Linear trend model

FIGURE 12 Mean bias of the treatment effect for settings with the A, step function and B, the linear trend function. The depicted mean biases were calculated by including all scenarios with the same interaction effect and the same sample sizes

A well known problem after identifying a subgroup is that the treatment effect in this identified subgroup is likely to be overestimated. Figure 12 shows the biases of the naive estimator of the treatment difference $\hat{z}(\widehat{BM+})$ for the five different subgroup identification methods. We calculated the depicted biases with $\widehat{\text{bias}} = \frac{1}{D}\sum_{i=1}^{D} \hat{z}(\widehat{BM+})^{(i)} - z(\widehat{BM+})^{(i)}$ where D denotes the number of generated datasets with differing prognostic effect and cut-off value but the same combination of sample size, interaction effect size, and mean function. The estimated and true treatment effect in the identified $\widehat{BM+}$ of dataset i is denoted $\hat{z}(\widehat{BM+})^{(i)}$ and $z(\widehat{BM+})^{(i)}$, respectively. Note that we did not include cases where the overall population was chosen to be the $\widehat{BM+}$ or $\widehat{BM-}$ group in the calculation of $\widehat{\text{bias}}$. Therefore, the depicted averages are based on different numbers of observations (see the Type II error of the methods, eg, Figure 6). In general, we can observe that the biases are larger when the treatment effect in the true underlying subgroup is small. Moreover, the biases tend to get larger with decreasing sample size. As seen in the previous simulations results regarding the selection accuracy, the chances of identifying the right $BM+$ subgroup increase with increasing effect in the true $BM+$ subgroups and with increasing sample sizes. For SIDES, we can observe the smallest values for the biases. But we have to keep in mind that the $\widehat{BM+}$ subgroups identified by SIDES include only few observations less than the overall study population. As expected, we overestimate the true treatment effect of the identified subgroups. An exception of a positive bias can be observed for subgroups identified by ARDP in settings with the step function, when the treatment effect in the true subgroup is large. Only in settings with $c = 0.5$, a $\widehat{BM+}$ subgroup different from the overall population is selected in some simulation runs. This is based on the lower true treatment effect in the overall population compared with the considered settings with equally sized $BM+$ and $BM-$ subgroups ($c = 0$) and a $BM+$ subgroup larger than its complement ($c = -0.5$).

6 | DISCUSSION

The range of methods developed for identifying subgroups with a homogeneous treatment effect is wide. However, little has been done so far to evaluate their performances in comparison based on realistic trial scenarios and using criteria appropriate for drug development. We chose five methods proposed for subgroup identification in order to select a future target population and compared them numerically by applying them to a data set from ALS and by means of simulations. Therefore, we selected a target population from the results obtained by MOB, IT, SIDES, and STIMA. For this selection, we combined those identified subgroups with a treatment effect considered as relevant. For ARDP, which results in a sequence of possible target populations, we additionally imposed that the final subgroup should be the largest subgroup showing a relevant treatment effect. We used the raw, nonstandardized treatment effect in the subgroup criterion as clinical relevance is not just indicated by the effect size but also by the outcome itself.[42] Nevertheless, imposing a threshold for the standardized treatment effect could also be justified. The selection of potential future target populations based on treatment effects is of special interest in a regulatory context. Therefore, our comparative study focuses on a selection criterion appropriate from a regulatory point of view. Other comparisons are based on a single dataset,[21] evaluate the estimation of tree-based treatment regimes[23] or focus on identifying covariates affecting treatment effects[22] rather than identifying subgroups.

For the authorization of a new drug, the key importance is showing a positive risk/benefit balance in a specific target population used in the indication. Therefore, the study population in which this positive balance is indented to be demonstrated should “closely mirror the target population.”[43] In some scenarios as described by the subgroup draft guideline,[44] a restriction of the target population might be of interest in order to increase the chances of a successful following trial. Usually, those restrictions are based on expected treatment heterogeneity. IT, STIMA, and MOB enable the user to find such a restriction of the target population with a low type I error rate. When a new trial is then planned in a restricted target population, some justification should be presented for the exclusion of certain patients. A justification purely relying on the results of such a method might not be accepted by regulatory authorities. Adding a biological rational referring to the biological mechanism of a drug to the exploratory subgroup findings can help to evaluate the credibility of subgroup findings. Note that the biological plausibility should imperatively focus on the mechanism of action, since all subgroups defined by any combination of prognostic biomarkers seem biologically reasonable.[45]

On the basis of the simulation study conducted, we found that MOB, STIMA, and IT can be used in settings with larger sample sizes as all three methods perform comparably well in terms of the proportion of correctly classified patients. Although STIMA showed a slightly better performance compared with IT throughout the scenarios considered, IT has the advantage of a more intuitive interpretation of its result with regard to the assignment rules. However, MOB exhibits the best performance in most of the settings. The exception are settings where biomarkers have a linear prognostic effect although they are not predictive. Seibold et al[46] addressed this setting with a proposed extension of MOB called PALM trees. The effect of the prognostic variables are assumed to be constant over all subgroups as it was the case in our simulation study. Furthermore, variables being only prognostic but not predictive have to be known for the use of PALM trees. If purely prognostic variables are not known in advance and are used as splitting candidates STIMA leads to the best results. Nevertheless, STIMA needs larger sample sizes or larger treatment effect differences in the target population and its complement to achieve satisfactory performance. Seibold et al[16] suggested to examine the estimated model parameters in the partitions obtained by MOB in order to decide whether splitting variables are prognostic only. A constant treatment effect in nodes with the same parent-node suggests covariates being purely prognostic.

In the simulations SIDES' performance is not convincing when an overall treatment effect is present. Since the splitting criterion of SIDES seems to evaluate the difference in precision between two nodes resulting from a split, SIDES tends to identify higher proportions of spurious subgroups when larger overall treatment effects and larger sample sizes are present.[47,48] Therefore, Mistry et al[48] proposed a new splitting criterion for SIDES considering the differential effect of two nodes rather than the precision. ARDP's performance in the simulations, however, is not convincing mainly due to the criterion used to select the final subgroup. Although, the performance of MOB is good in many of the considered settings, it still performs poorly in settings with smaller sample sizes. All considered methods do not show a convincing classification for the smaller sample size of our simulation study. But the identification of a future target population is of special interest after early phase trials, which usually involve much less than 600 patients. This problem can be addressed by pooling data from different trials on the same drug, which should not be done naively. Pooling the data should be done by an individual-patient meta-analysis framework allowing to account for heterogeneity. Examples were proposed by Patel et al,[19] Mistry et al,[48] and Fokkema et al[49] who proposed extensions to some of the here mentioned subgroup identification methods. However, systematic evaluation and practical experience of these extensions are lacking in particular with the view on regulatory settings.

ACKNOWLEDGEMENTS

This work was cofunded by the BfArM and the University Medical Center Göttingen and is a result of the collaborative project “Identifikation und Konfirmation von Biomarker-definierten Populationen in der personalisierten Pharmakotherapie (“Identification and confirmation of biomarker-defined populations in the personalized pharmacotherapy”).

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the Pooled Resource Open-Access ALS Clinical Trials (PRO-ACT) Database. Restrictions apply to the availability of these data, which were used under licence for this study. Data are available at https://nctu.partners.org/ProACT/ with the permission of the PRO-ACT Consortium.

ORCID

Cynthia Huber https://orcid.org/0000-0003-2035-3682
Norbert Benda https://orcid.org/0000-0001-5605-2414
Tim Friede https://orcid.org/0000-0001-5347-7441

REFERENCES

1. Huber C, Friede T, Stingl J, Benda N. Classification of biomarkers to be used as companion diagnosticy: consequences for regulatory decisions on biomarker driven patient selection. 2018. Manuscript in preparation.
2. Biomarkers Definitions Working Group. Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. *Clin Pharmacol Ther*. 2001;69(3):89-95.
3. Amado RG, Wolf M, Peeters M, et al. Wild-type KRAS is required for panitumumab efficacy in patients with metastatic colorectal cancer. *J Clin Oncol*. 2008;26(10):1626-1634.
4. Mandrekar SJ, Sargent JD. Clinical trial designs for predictive biomarker validation: theoretical considerations and practical challenges. *J Clin Oncol*. 2009;27(24):4027-4034.
5. De Roock W, Piessevaux H, De Schutter J, et al. KRAS wild-type state predicts survival and is associated to early radiological response in metastatic colorectal cancer treated with cetuximab. *Ann Onco*. 2008;19(3):508-515.
6. Di Fiore F, Blanchard F, Charbonnier F, et al. Clinical relevance of KRAS mutation detection in metastatic colorectal cancer treated by Cetuximab plus chemotherapy. *Brit J Cancer*. 2007;96(8):1166-1169.
7. Lièvre A, Bachet JB, Le Corre D, et al. KRAS mutation status is predictive of response to cetuximab therapy in colorectal cancer. *Cancer Res*. 2006;66(8):3992-3995.
8. Brannath W, Zuber E, et al. Confirmatory adaptive designs with bayesian decision tools for a targeted therapy in oncology. *Stat Med*. 2009;28(10):1445-1463.
9. Jenkins M, Stone A, Jennison C. An adaptive seamless phase ii/iii design for oncology trials with subpopulation selection using correlated survival endpoints. *Pharm Stat*. 2011;10(4):347-356.
10. Friede T, Parsons N, Stallard N. A conditional error function approach for subgroup selection in adaptive clinical trials. *Stat Med*. 2012;31(30):4309-4320.
11. Wang S-J. Utility of adaptive strategy and adaptive design for biomarker-facilitated patient selection in pharmacogenomic or pharmacogenetic clinical development program. *J Formosan Med Assoc*. 2008;107(12, Supplement):S19-S27.

12. Bauer P, Bretz F, Dragalin V, König F, G Wassmer. Twentyfive years of confirmatory adaptive designs: opportunities and pitfalls. *Stat Med*. 2015;35(3):325-347.
13. Ondra T, Dmitrienko A, Friede T, et al. Methods for identification and confirmation of targeted subgroups in clinical trials: a systematic review. *J Biopharm Stat*. 2016;26(1):99-119.
14. Lipkovich I, Dmitrienko A, D'Agostino RB. Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Stat Med*. 2017;36(1):136-196.
15. Su X, Tsai C-L, Wang H, Nickerson DM, Li B. Subgroup analysis via recursive partitioning. *J Mach Learn Res*. 2009;10:141-158.
16. Seibold H, Zeileis A, Hothorn T. Model-based recursive partitioning for subgroup analyses. *Int J Biostat*. 2016;12(1):45-63.
17. Lipkovich I, Dmitrienko A, Denne J, Enas G. Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat Med*. 2011;30(21):2601-2621.
18. Dusseldorp E, Conversano C, Van Os BJ. Combining an additive and tree-based regression model simultaneously: Stima. *J Comput Graph Stat*. 2010;19(3):514-530.
19. Patel S, Hee SW, Mistry D, et al. Identifying back pain subgroups: developing and applying approaches using individual patient data collected within clinical trials. *Programme Grants Appl Res*. 2016;4(10):1-314.
20. Boulesteix A-L, Binder H, Abrahamowicz M, Sauerbrei W. On the necessity and design of studies comparing statistical methods. *Biometrical J*. 2018;60(1):216-218.
21. Doove LL, Dusseldorp E, Van Deun K, Van Mechelen I. A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment—subgroup interactions. *Adv Data Anal Classi*. 2014;8(4):403-425.
22. Alemayehu D, Chen Y, Markatou M. A comparative study of subgroup identification methods for differential treatment effect: performance metrics and recommendations. *Stat Methods Med Res*. 2018;27(12):3658-3678.
23. Sies A, Van Mechelen I. Comparing four methods for estimating tree-based treatment regimes. *Int J Biostat*. 2017;13:1-23.
24. Benda N, Branson M, Maurer W, Friede T. Aspects of modernizing drug development using clinical scenario Planning and Evaluation. *Drug Inf J*. 2010;44(3):299-315.
25. EMA. Concept paper on predictive biomarker-based assay development in the context of drug development and lifecycle. European Medicines Agency. EMA/CHMP/800914/2016; 2017.
26. Li M, Yu T, Hu Y-F. The impact of companion diagnostic device measurement performance on clinical validation of personalized medicine. *Stat Med*. 2015;34(14):2222-2234.
27. Breiman L, Friedman JH, Olshen RA, Stone CJ. *Classification and Regression Trees*. Monterey,CA: Wadsworth and Brooks; 1984.
28. LeBlanc M, Crowley J. Survival trees by goodness of split. *J Am Stat Assoc*. 1993;88(422):457-467.
29. Zeileis A, Hothorn T, Hornik K. Model-based recursive partitioning. *J Comput Graph Stat*. 2008;17(2):492-514.
30. LeBlanc M, Moon J, Crowley J. Adaptive risk group refinement. *Biometrics*. 2005;61(2):370-378.
31. Zhao L, Tian L, Cai T, Claggett B, Wei LJ. Effectively selecting a target population for a future comparative study. *J Am Stat Assoc*. 2013;108(502):527-539.
32. Brown RH, Al-Chalabi A. Amyotrophic lateral sclerosis. *New Engl J Med*. 2017;3:17071.
33. FDA. Radicava (edaravone injection), for intravenous use. U.S. Food and Drug Administration. Reference ID: 4094543; 2017.
34. FDA. Labeling: Rilutek (riluzole) tablets. U.S. Food and Drug Administration. Application Number NDA 20599 S-011 & S-012; 2009.
35. Atassi N, Berry J, Shui A, et al. The PRO-ACT database design, initial analyses, and predictive features. *Neurology*. 2014;83(19):1719-1725.
36. Prize4Life Israel and Neurological Clinical Research Institute. Pooled resource open-access als clinical trials database. Massachusetts General Hospital. https://nctu.partners.org/ProACT/; 2015.
37. Chio A, Lagroscino G, Hardiman O, et al. Prognostic factors in als: a critical review. *Amyotrophic Lateral Sclerosis*. 2009;10:310-323.
38. Hothorn T, Zeileis A. partykit: a modular toolkit for recursive partytioning in R. *J Mach Learn Res*. 2015;16:3905-3909.
39. Dusseldorp E, Conversano C. Stima: simultaneous threshold interaction modeling algorithm. R package version 1.1; 2013.
40. Graf AC, Wassmer G, Friede T, Gera RG, Posch M. Robustness of testing procedures for confirmatory subpopulation analyses based on a continuous biomarker. *Stat Methods Med Res*. 2018;0(0):0962280218777538.
41. Friede T, Henderson R. Exploring changes in treatment effects across design stages in adaptive trials. *Pharm Stat*. 2009;8(1):62-72.
42. Cuijpers P, Turner EH, Koole SL, Dijke A, Smit F. What is the threshold for a clinically relevant effect? the case of major depressive disorders. *Depress Anxiety*. 2016;31(5):374-378.
43. ICH. Statistical principles for clinical trials. ICH E9 Expert Working Group. CPMP/ICH/363/96; 1998.
44. EMA. Guideline on the investigation of subgroups in confirmatory trials (draft). European Medicines Agency/Committee for Medical Products for Human Use. EMA/CHMP/539146/2013; 2014.
45. Dmitrienko A, Muysers C, Fritsch A, Lipkovich I. General guidance on exploratory and confirmatory subgroup analysis in late-stage clinical trials. *J Biopharm Stat*. 2016;26(1):71-98.
46. Seibold H, Hothorn T, Zeileis A. Generalised linear model trees with global additive effects. *Adv Data Anal Classif*. 2018:1-23.
47. Mistry D. Recursive partitioning based approaches for low back pain subgroup identification in individual patient data metaanalyses. 2014.
48. Mistry D, Stallard N, Underwood M. A recursive partitioning approach for subgroup identification in individual patient data metaanalysis. *Stat Med*. 2018;37(9):1550-1561.
49. Fokkema M, Smits N, Zeileis A, Hothorn T, Kelderman H. Detecting treatment-subgroup interactions in clustered data with generalized linear mixed-effects model trees. *Behav Res Methods*. 2018;50(5):2016-2034.

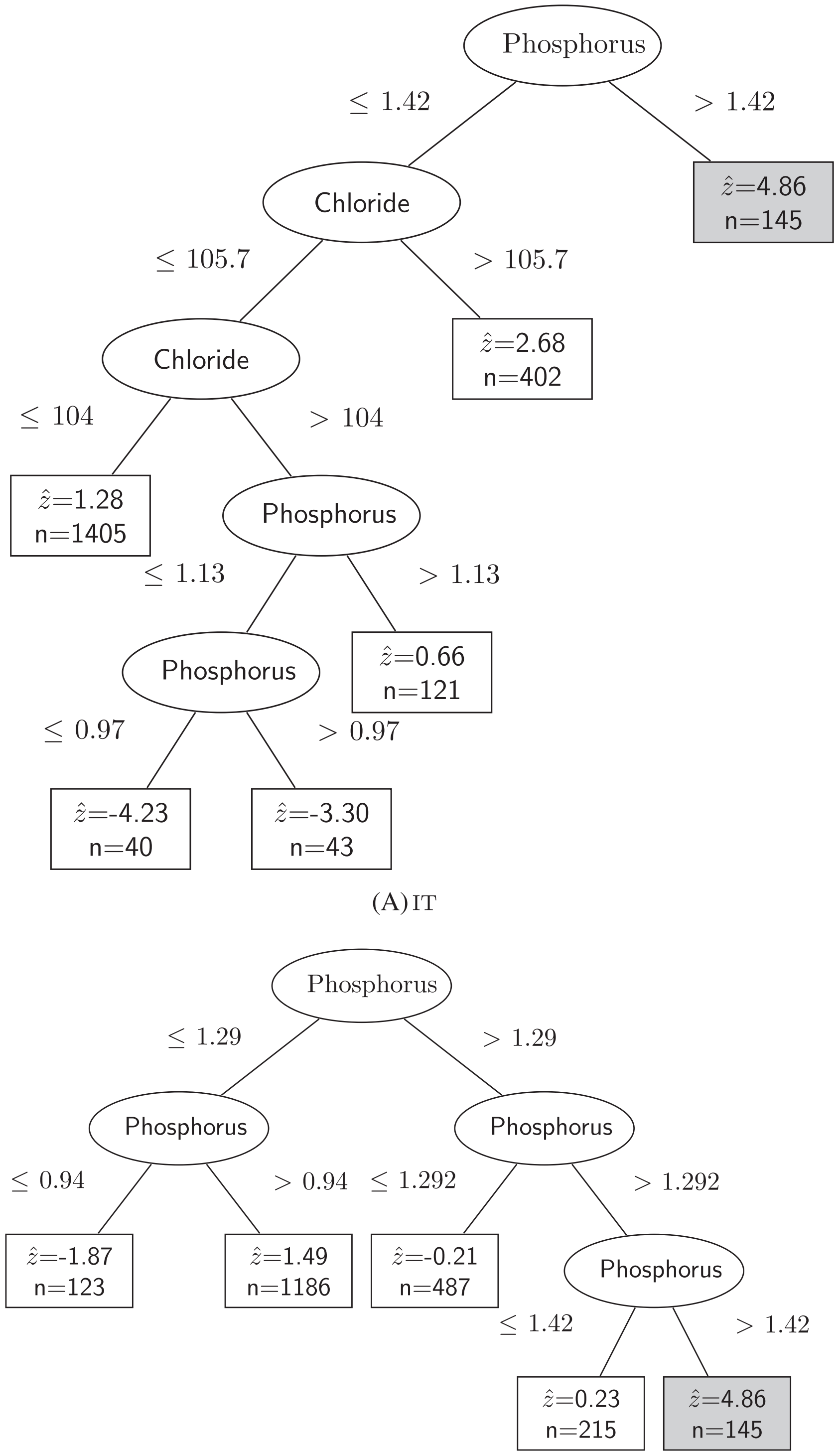
How to cite this article: Huber C, Benda N, Friede T. A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations. *Pharmaceutical Statistics*. 2019;18:600–626. https://doi.org/10.1002/pst.1951

APPENDIX A: GROWN TREES OBTAINED BY APPLYING THE METHODS TO ALS DATA

In the following, the trees which are grown by IT, MOB, STIMA, SIDES, and ARDP are illustrated. For IT and STIMA, no *BM*+ subgroup is selected when the originally proposed pruning procedures are used. Therefore, the initial trees of those two procedures are illustrated in Figures A1 and A2.

For IT and MOB, the size and the treatment difference z in each identified subgroup are given in Figure A1. The leafs fulfilling the subgroup criterion are colored in gray. IT grows an initial tree with one split on the variable chloride resulting in two subgroups of unequal size. The subgroup involving 131 patients has an estimated treatment effect of 5.3 and is therefore assigned as *BM*+ subgroup. An initial tree involving just one split is a comparable small tree. Here, this small initial tree is a result of the chosen tuning parameter for the minimum size of an end node. As IT, MOB also uses the estimated treatment difference $\hat{z}$ to decide which leave is considered belonging to the *BM*+ or *BM*− subgroup.

For the STIMA solution, the mean outcome was calculated in each leaf. In order to define the *BM*+ group, further steps are necessary (see Section 3). The results of SIDES and ARDP are shown in Figure A3. SIDES does not identify any candidate subgroups. The estimated $\widehat{BM+}$ subgroup using ARDP is colored in gray. The treatment differences in the sequence of subgroups is denoted by $\hat{z}$. The illustrated solution of ARDP does not include all estimated peeling steps.



(A) IT

(B) MOB

FIGURE A1 Trees obtained by A, interaction trees and B, model-based recursive partitioning

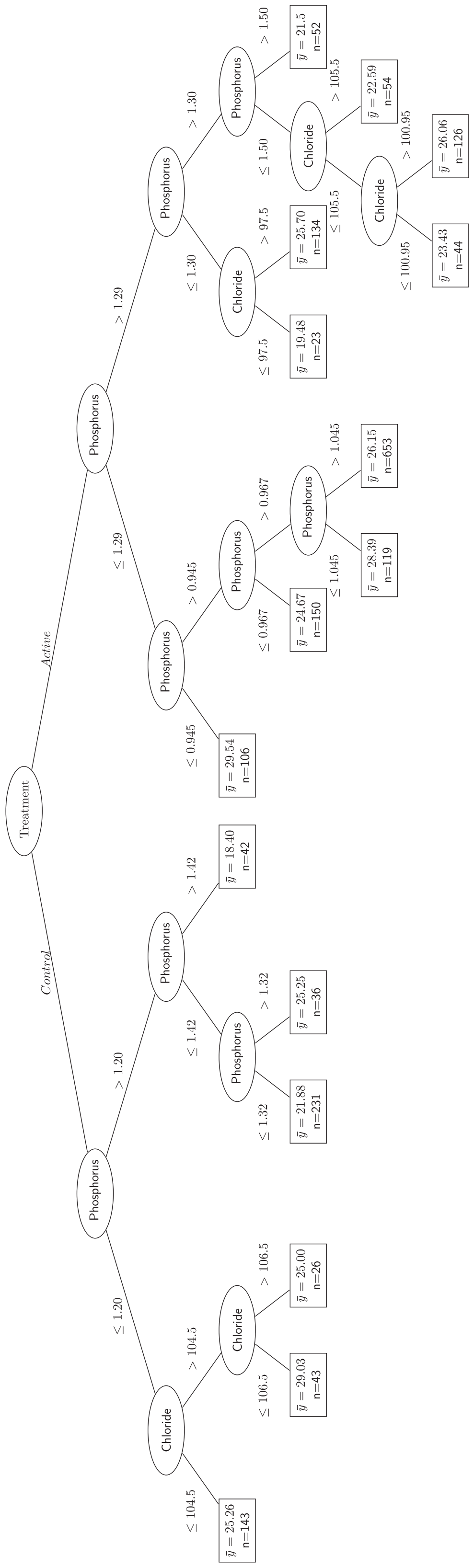
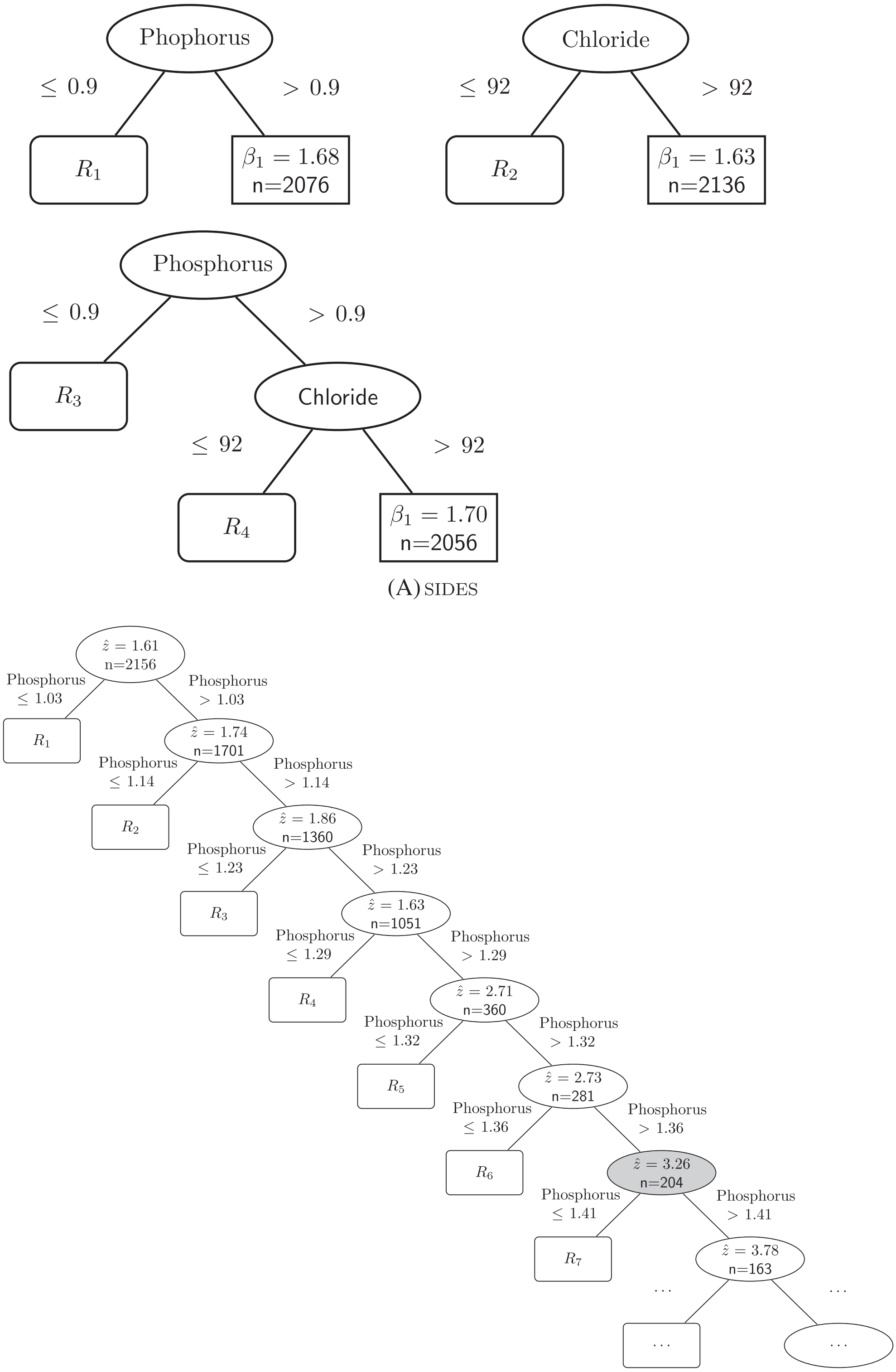


FIGURE 14 Tree obtained by simultaneous threshold interaction modeling algorithm (STIMA)



(A) SIDES

(B) ARDP

FIGURE 15 A, Subgroup identification based on differential effect search (SIDES) does not identify any subgroup fulfilling the subgroup criterion presented in Section 3. B, The illustrated result of adaptive refinement by directed peeling algorithm (ARDP) does not involve all iteration steps