Detecting Treatment-Subgroup Interactions in Clustered Data with Generalized Linear Mixed-Effects Model Trees

M. Fokkema[1], N. Smits[2], A. Zeileis[3], T. Hothorn[4], H. Kelderman[5]
[1]Universiteit Leiden, [2]Universiteit van Amsterdam, [3]Universität Innsbruck, [4]Universität Zürich, [5]Universiteit Leiden and Vrije Universiteit, Amsterdam

Abstract

Identification of subgroups of patients for which treatment A is more effective than treatment B, and vice versa, is of key importance to the development of personalized medicine. Several tree-based algorithms have been developed for the detection of such treatment-subgroup interactions. In many instances, however, datasets may have a clustered structure, where observations are clustered within, for example, research centers, studies or persons. In the current paper we propose a new algorithm, generalized linear mixed-effects model (GLMM) trees, that allows for detection of treatment-subgroup interactions, as well as estimation of cluster-specific random effects. The algorithm uses model-based recursive partitioning (MOB) to detect treatment-subgroup interactions, and GLMMs for estimation of random-effects parameters. In a simulation study, we evaluate the performance of GLMM trees and compare it with that of MOB trees without random-effects estimation. In datasets without treatment-subgroup interactions, GLMM tree was found to have a much lower Type I error rate than MOB trees without random effects (4% and 33%, respectively). Furthermore, in datasets with treatment-subgroup interactions, GLMM trees recovered the true treatment subgroups much more often than MOB without random effects (90% and 61% of the datasets, respectively). Also, GLMM trees predicted treatment outcome differences more accurately than MOB trees without random effects (average accuracy of .94 and .88, respectively). We illustrate the application of GLMM tree on a patient-level dataset of a meta-analysis on the effects of psycho- and

The authors would like to thank Prof. Pim Cuijpers, Prof. Jeanne Miranda, Dr. Boadie Dunlop, Prof. Rob DeRubeis, Prof. Zindel Segal, Dr. Sona Dimidjian, Prof. Steve Hollon and Erica Weitz for granting access to the dataset for the application. The work for this paper was partially done while MF, AZ and TH were visiting the Institute for Mathematical Sciences, National University of Singapore in 2014. The visit was supported by the Institute.

pharmacotherapy for depression. We conclude that GLMM trees are a promising algorithm for the detection of treatment-subgroup interactions in clustered datasets, and discuss some directions for future research.

Keywords: model-based recursive partitioning, treatment-subgroup interactions, random effects, generalized linear mixed-effects model, classification and regression trees

Introduction

In research assessing the efficacy of treatments for somatic and psychological disorders, the one-size-fits-all paradigm is slowly losing ground, and stratified medicine is becoming increasingly important. Stratified medicine presents the challenge of finding which patients respond best to which treatments. This can be referred to as the detection of treatment-subgroup interactions (e.g., Doove, Dusseldorp, Van Deun, & Van Mechelen, 2014). In most cases, treatment-subgroup interactions are studied using linear models, such as factorial analysis of variance techniques, in which potential moderators have to be specified a-priori, have to be checked one at a time, and continuous moderator variables have to be discretized a-priori. This may hamper identification of which treatments work best for whom, especially when there are no a-priori hypotheses about treatment-subgroup interactions. As noted by Kraemer, Frank, and Kupfer (2006), there is a need for methods that generate, instead of test, hypotheses and that are specifically directed at the detection of treatment interactions.

Tree-based methods are such hypothesis-generating methods, as they can automatically detect subgroups which differ in the expected outcomes for one or more treatments. Due to their flexibility, tree-based methods are preeminently suited to the detection of treatment-subgroup interactions: they can handle many potential predictor variables at once and can automatically detect (higher order) interactions between predictor variables (Strobl, Malley, & Tutz, 2009). Several promising tree-based algorithms for the detection of treatment-subgroup interactions have been developed (e.g., Dusseldorp & Van Mechelen, 2014; Dusseldorp & Meulman, 2004; Su, Tsai, Wang, Nickerson, & Li, 2009; Foster, Taylor, & Ruberg, 2011; Lipkovich, Dmitrienko, Denne, & Enas, 2011; Zeileis, Hothorn, & Hornik, 2008; see Doove et al., 2014 for an overview). Among these methods, model-based recursive partitioning (MOB; Zeileis et al., 2008) seems to be the most flexible tool for detecting treatment-subgroup interactions as it offers a generic inferential framework that can be coupled with a broad range of parametric modeling strategies fitted by M-type estimators (Zeileis et al., 2008). Specifically, this model class encompasses the generalized linear model (GLM) and GLM-based MOB trees have been successfully applied by Driessen et al. (2014) in the detection of subgroups with differential treatment outcomes for two different psychotherapies.

However, none of the aforementioned tree-based algorithms allow for taking into account the clustered structure of many datasets. In such cases, researchers may want to detect treatment-subgroup interactions in datasets with a clustered structure (e.g., Koopman, Van der Heijden, Glasziou, Grobbee, & Rovers, 2007). For example, in individual-level patient data meta-analyses, in which datasets of multiple trials evaluating the effects of the same treatments are pooled. In such analyses, the clustered structure of the dataset should be taken into account by including study-specific effects in the model, prompting the need for modeling random effects (e.g., Cooper & Patall, 2009; Higgins, Whitehead, Turner, Omar, & Thompson, 2001). Likewise, longitudinal datasets, and datasets from multi-center trials typically also require modeling of random effects. Ignoring the clustered structure of datasets may lead to biased inference due to underestimated standard errors (e.g., Bryk & Raudenbush, 1992; Van den Noortgate, Opdenakker, & Onghena, 2005). More specifically, when the interest is in subgroup detection, ignoring random effects on the outcome variable may result in the detection of spurious subgroups (e.g., Sela & Simonoff, 2012).

In the current paper, we present a tree-based algorithm for detecting treatment-subgroup interactions, which takes the clustered nature of datasets into account. The algorithm combines MOB with random-effects estimates and therefore accounts for both the clustering structure (which is not done in other tree-based treatment-subgroup interaction detection methods, e.g., Zeileis et al., 2008; Su et al., 2009; Dusseldorp & Van Mechelen, 2014) and the treatment effect estimation in models with continuous and non-continuous response variables (which is not available in previously suggested regression trees with random effects, e.g., Hajjem, Bellavance, & Larocque, 2011; Sela & Simonoff, 2012).

In what follows, we will introduce the existing frameworks for estimating treatment effects: the generalized linear model (GLM), model-based recursive partitioning (MOB), and the generalized linear mixed-effects model (GLMM). Then, we introduce our new algorithm, which combines MOB and the GLMM: generalized linear mixed-effects model trees. Subsequently, we evaluate the performances of the GLMM tree algorithm in simulated datasets, and illustrate application of the algorithm with an exisiting dataset of a patient-level meta-analysis on the effects of treatments for depression by Cuijpers et al. (2014). Before we discuss the methods for estimating treatment effects, we will introduce an artificial motivating data set, with which the methods will be illustrated. After we introduce the GLMM tree algorithm, we present a simulation study, in which we evaluate GLMM trees comparative accuracy. Finally, in the application, we use GLMM tree to detect treatment-subgroup interactions in an existing dataset on the effects of treatments for depression.

Artificial motivating dataset

To illustrate the application of the methods to be discussed, we will use a simulated dataset of 150 observations, which were randomly assigned to Treatment 1 or Treatment 2.

Every observation has a value for the response variable, with which the effect of treatment is assessed: the posttreatment total score on a depression inventory. Further, all observations have values for three covariates: duration of depressive symptoms prior to treatment in months (range 0–15); age in years (range 18–75); anxiety inventory total score (range 3–18).

The simulated dataset has 3 subgroups with different treatment effectiveness. The first subgroup consists of observations with duration ≤ 6 and anxiety ≤ 10. In this subgroup, Treatment 1 is more effective than Treatment 2: the mean of the response variable is 7 for Treatment 1, and 11 for Treatment 2. The second subgroup consists of observations with duration ≤ 6 and anxiety > 10. In this subgroup, both therapies are equally effective: the mean value of the response variable is 9 for Treatment 1, and 9 for Treatment 2. The third subgroup consists of observations with duration > 6. In this subgroup, Treatment 2 is more effective than Treatment 1: the mean value of the response variable is 12 for Treatment 1, and 7 for Treatment 2.

Observations were drawn from one of ten clusters, each with a different, cluster-specific (i.e., random) intercept. Data was generated such that covariates and cluster-specific intercepts were uncorrelated. Also, 43% of variance in posttreatment depression scores was due to treatment-subgroup interactions, and 8% of variance was due to cluster-specific variation.

General modeling framework

GLM

In a clinical trial, where the outcomes of two or more treatments are compared, an overall GLM is often used to estimate treatment effects. GLMs allow for the choice of a suitable response distribution – for example normal, binomial, or Poisson - depending on whether the treatment outcome variable is continuous (e.g., an improvement score), binary (e.g., improved or not), or a count (e.g., number of events in a certain time span), respectively. In all cases the expectation μ_i of the outcome variable y_i given the treatment regressors x_i is modeled through a linear predictor and a suitable link function:

$$E[y_i|x_i] = \mu_i, \tag{1}$$

$$g(\mu_i) = x_i^\top \beta, \tag{2}$$

where $x_i^\top \beta$ is the linear predictor for observation i and g is the link function[1]. Further, x_i is a vector of fixed-effects predictor variable values for observation i, of which the first element takes a value of 1 for the intercept, and the second element takes the value of a dummy

[1] An overview of notation used is provided in the appendix.

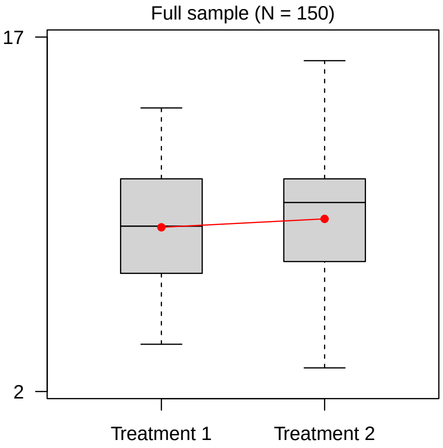

Figure 1. Example of a normal GLM (with fixed effects only) for treatment outcomes, based on the artificial motivating dataset ($N = 150$). The dot for Treatment 1 represents the first, and the slope of the regression line represents the second element of β.

indicator for treatment type (a value of 0 for the first, or reference treatment type, and a value of 1 for the second, or focal treatment type). β is a vector of fixed-effects regression coefficients, the first element representing the intercept, which is the mean value of the linear predictor in the first treatment group, and the second element representing the slope, which is the mean difference in the linear predictor between the first and second treatment groups. In case of a continuous response variable, we employ a Gaussian distribution with identity link and denote the error by $\epsilon_i = y_i - \mu_i$ with variance σ^2_ϵ.

To keep notation and examples simple, we assume x_i and β to have length 2. That is, the effects of only two treatment conditions are estimated and no additional covariates are included in the GLM. However, additional treatment conditions and covariates can easily be included. In addition, examples and datasets in the current paper will focus on continuous response variables with normally distributed errors, such as posttreatment severity of a disorder. But the models and algorithms to be discussed can also be applied with discrete outcomes, such as remission of a disorder (yes/no).

To illustrate, the GLM estimated for the artificial motivating dataset is graphically represented in Figure 1. The boxplots in Figure 1 show the distribution of the posttreatment depression scores in both treatment groups. There seems to be little overall difference in effects of both treatments, as the slope of the regression line is nearly zero. We shall see that this does not necessarily mean that posttreatment depression score and treatment type are unrelated, as the effect of treatment may be moderated by variables not yet included in the model.

Model-based recursive partitioning

The rationale behind MOB is that a global model for all observations, like the GLM in Equation 1 and 2, may not describe all data well, and when additional covariates are available it may be possible to partition the dataset with respect to these covariates, and find a better model in each cell of the partition (Zeileis et al., 2008). This is reminiscent of the classification and regression tree (CART) algorithm of Breiman, Friedman, Olshen, and Stone (1984), which splits the dataset into subsets, for which the distributions of the outcome variable are most different. However, CART trees detect differences in constant fits across terminal nodes, whereas MOB trees detect differences in parametric models across terminal nodes.

To find partitions and better-fitting local GLMs, the MOB algorithm tests for parameter instability. When the partitioning is based on a GLM, instabilities are differences in $\hat{\beta}$ across partitions of the dataset, which are defined by one or more auxiliary covariates not included in the linear predictor. To find partitions, the MOB algorithm cycles iteratively through the following steps (Zeileis et al., 2008): (1) fit the parametric model to the dataset, (2) test for parameter instability over a set of partitioning variables, (3) if there is some overall parameter instability, split the dataset with respect to the variable associated with the highest instability, (4) repeat the procedure in each of the resulting subgroups.

More specifically, in step (2), to test for parameter instability, the so-called *scores* are computed, using the score function. By definition, the empirical scores of all observations in a dataset sum to zero, and when the model is correctly specified, the expected value of the score for each observation is also zero. Under the null hypothesis of parameter stability, the scores do not systematically deviate from the expected value of zero, when the observations are ordered by the values of a potential partitioning variable U_k (c.f., Merkle & Zeileis, 2013). To statistically test whether the scores systematically deviate from zero with respect to variable U_k, the class of generalized M-fluctuation tests is used (Zeileis, 2005; Zeileis & Hornik, 2007).

If the null hypothesis of parameter stability in step (2) can be rejected, that is, if at least one of the partitioning variables U_k has a p-value for the M-fluctuation test below the pre-specified significance level α, the dataset is partitioned into two subsets in step (3). In step (3), a binary partition is created using U_{k*}, the variable with the minimal p-value in step (2). The split point for U_{k*} is selected, by taking the value that minimizes the sum of the values of the objective function in both partitions (Zeileis et al., 2008). In step (4), steps (1) through (3) are repeated in each partition, until the null hypothesis of parameter stability can no longer be rejected.

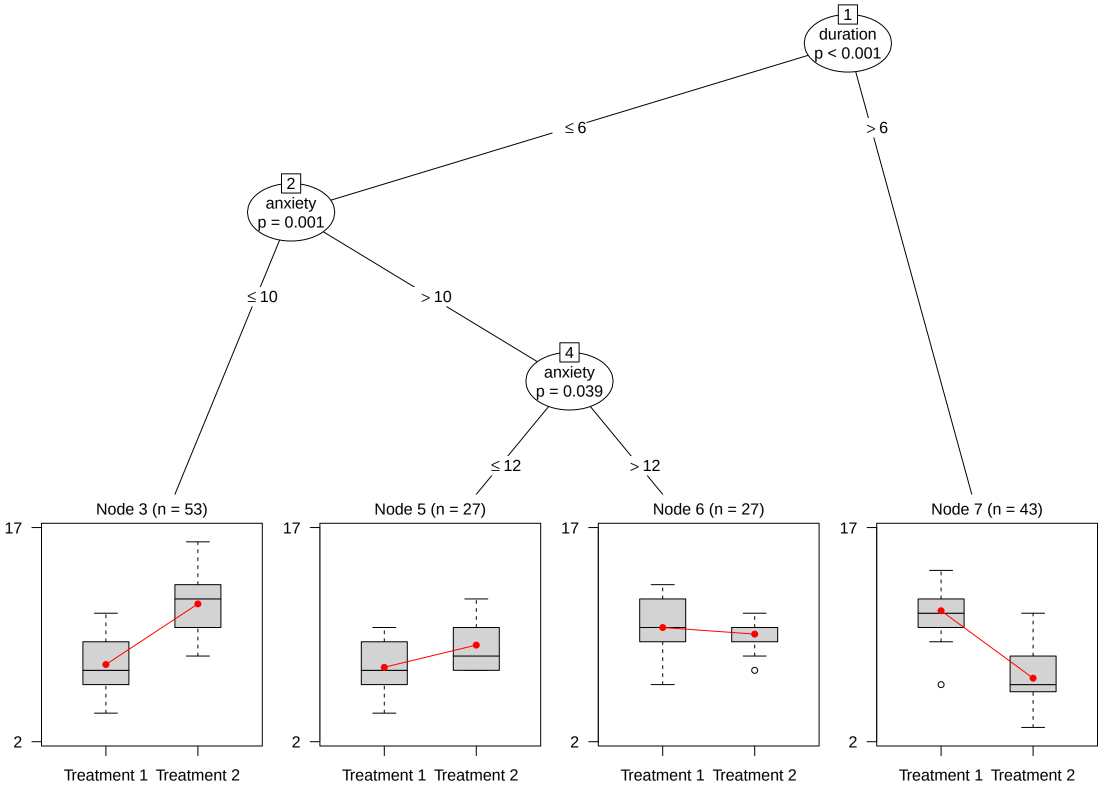

Figure 2. Example of a tree representation of model-based recursive partition, based on the artificial motivating dataset. Three additional covariates (anxiety questionnaire score, duration of depressive symptoms at baseline in months and age) were used as potential splitting variables.

Due to the binary recursive nature of MOB, the resulting partition can be represented as a binary tree. If the partitioning is based on the GLM, the result is a GLM tree, which has a local fixed-effects regression model in every j-th ($j = 1, \ldots, J$) terminal node of the tree. As a result, in the GLM tree model, the value for β depends on terminal node j in which observation i 'falls':

$$g(\mu_{ij}) = x_i^\top \beta_j \tag{3}$$

Note that, if the recursive subgroup structure (i.e., the partition) were known, the tree could be estimated as a single GLM where all coefficients interact with the factor indicating the subgroup. Somewhat more formally, the model could then be written: $g(\mu_i) = x_i^{*\top} \beta^*$, where x_i^* are the values of the $2J$ interactions between the subgroups from the tree, and the elements of x_i. β^* would also have length $2J$, and contain the subgroup-specific fixed-effects coefficients.

Figure 2 provides an example of the GLM tree model in Equation 3, based on the artificial motivating dataset. By using the three additional covariates (anxiety, duration and age), MOB partitioned the observations into four subgroups, each with a different estimate for β_j. Age was correctly not detected as a partitioning variable, and the left- and rightmost subgroups are in accordance with the treatment-subgroup interactions as described above. However, the two subgroups in the middle result from a spurious split.

GLMM

When a dataset contains observations from multiple clusters (e.g., trials, research centers, or individuals in longitudinal datasets), the GLM in Equation 2 may be extended to include cluster-specific, or random effects, and the model becomes a GLMM:

$$g(\mu_i) = x_i^\top \beta + z_i^\top b \tag{4}$$

Where z_i is a unit vector of length M, of which the m-th element takes a value of 1, and all other elements take a value of 0; m ($m = 1, \dots, M$) denotes the cluster which observation i is part of. Further, b is a random vector of length M, with every element being the random intercept for cluster m. Within the GLMM, it is assumed that b is normally distributed, with mean zero and variance σ_b^2. The parameters of the GLMM can be estimated with, for example, maximum likelihood (ML) and restricted ML (REML), as described in Bryk and Raudenbush (1992), for example.

For simplicity, we assume that only cluster-specific intercepts are included in the models. However, random-effects covariates and coefficients can easily be included.

Note that, if the random-effects coefficients were known, the model could be estimated by a simple GLM as in Equation 2 where $z_i^\top b$ would only be added as an offset (i.e., a variable with a fixed coefficient of 1) to the linear predictor.

GLMM tree

As noted earlier, ordinary GLM(M)s are not well suited for the detection of treatment-subgroup interactions, whereas the MOB algorithm is, but does not allow for estimation of random effects. Therefore, we propose the GLMM tree, which combines the GLMM from Equation 4 with the tree from Equation 3:

$$g(\mu_i) = x_i^\top \beta_j + z_i^\top b \tag{5}$$

To estimate the parameters of this model, we take an approach similar to that of Hajjem et al. (2011) and Sela and Simonoff (2012) but extend their ideas from classical CART trees with only random intercepts to a full GLMM algorithm. In the MERT approach, the fixed-effects part of a GLMM is replaced by a CART regression tree, and the random-effects

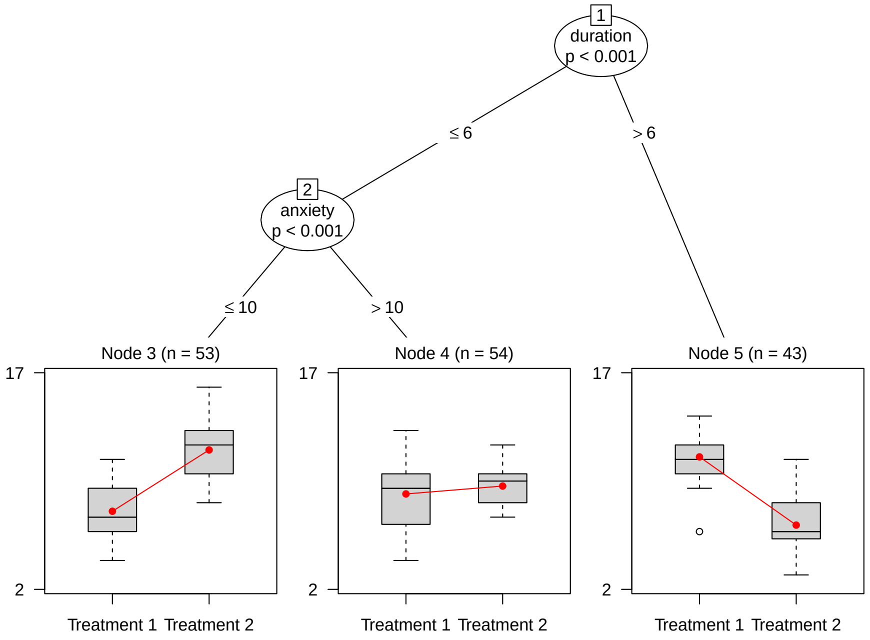

Figure 3. GLMM tree of the motivating example dataset. Three covariates (anxiety questionnaire score, duration of depressive symptoms at baseline in months and age) were used as potential splitting variables, and the clustering structure was taken into account by estimating random intercepts.

part is estimated as usual. To estimate a MERT, an iterative approach is taken, alternating between (1) assuming random effects known, allowing for estimation of the regression tree, and (2) assuming the regression tree known, allowing for estimation of the random effects.

For estimating GLMM trees, we take the MERT approach a step further: by using a GLM tree, instead of a CART regression tree with constant fits, to estimate the fixed-effects part of the GLMM. This allows not only for detection of differences in main effects, but also for detection of differences in regression effects (e.g., of treatment type) across terminal nodes. In addition, GLMM trees can be estimated for continuous, as well as binary and count variables. The GLMM tree algorithm takes the following steps to estimate the model in Equation 5:

Step 0: Initialize by setting r and all values $\hat{b}_{(r)}$ to 0.

Step 1: Set $r = r + 1$. Estimate GLM tree $(x_i^\top \hat{\beta}_{j(r)})$, with $z_i^\top \hat{b}_{(r-1)}$ as an offset.

Step 2: Estimate random effects in the mixed effects model $x_i^\top \hat{\beta}_{j(r)} + z_i^\top \hat{b}_{(r)}$ with subgroups $j(r)$ from the GLM tree.

Step 3: Repeat Steps 1 and 2 until convergence.

The algorithm initializes by setting all b values to 0, since the random-effects (and also the fixed-effects) parts are initially unknown. In every iteration, the GLM tree and random-effects coefficients b are re-estimated. The GLM tree is estimated, given the estimated $\hat{b}$ from the last iteration, and the b values are estimated, given the estimated GLM tree from the current iteration. Iterations are continued until convergence, which is monitored by computing the log-likelihood criterion of the mixed-effects model in Equation 4.

In Figure 3, the GLMM tree that was grown on the artificial motivating dataset is presented. As can be seen, by taking into account the clustering of observations by estimating random intercepts, the spurious split involving the anxiety variable no longer appears in the tree.

Emperical evaluation

We will asses the performance of GLMM tree in recovering treatment-subgroup interactions, and predicting differences between the outcomes of two treatments, in simulated datasets with continuous outcomes. In addition, we will compare the performance of GLMM tree with that of GLM tree. In the simulation study, the main interest will be in the effects of sample size, and the presence and magnitude of treatment-subgroup interactions and random effects, but other parameters will be varied, as well.

For GLMM tree, we expect the accuracy of recovered trees and predictions to improve with increasing sample size, and magnitude of the differences in treatment outcomes. For GLM tree, we have the same expectation, when random effects are absent; that is, when the variance of the random coefficients is zero, we expect GLM tree and GLMM tree to perform equally well. When random effects are present, we expect GLMM tree to perform better than GLM tree, and more so when the variance of random-effects coefficients is larger.

Simulation design

Datasets with treatment-subgroup interactions. For generating datasets with treatment-subgroup interactions, we used a treatment-subgroup interaction design from Dusseldorp and Van Mechelen (2014), which is also depicted in Figure 4. Figure 4 shows two subgroups with mean differences in treatment outcomes, and two subgroups without mean differences in treatment outcomes. The four subgroups are characterized by their values on the partitioning variables U_2, and U_1 or U_5. In other words, U_1, U_2 and U_5 are true partitioning variables, whereas the other potential partitioning variables (U_3, U_4, U_6 through U_{15}) are noise variables.

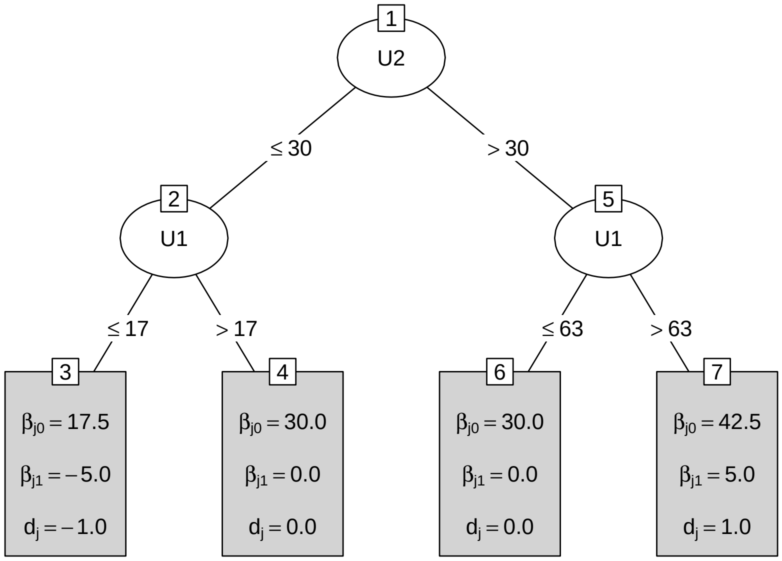

Figure 4. Data-generating model for treatment-subgroup interactions. Parameter d denotes the standardized mean difference between the outcomes of Treatment 1 and 2 (i.e., $\beta_{j1}/\sigma_{\epsilon}$).

Datasets without treatment-subgroup interactions. For generating datasets without treatment-subgroup interactions, we used a design in which there is only a main effect of treatment in the population. Put differently, the number of subgroups or terminal nodes in these datasets was $J = 1$, and there was only a single value of $\beta_j = \beta$ in every dataset. The mean of the outcome variable in the datasets without treatment-subgroup interactions was 30, which is the same value as in the datasets with treatment-subgroup interactions. As a result, $\beta = (27.5, 32.5)$ for all observations when $d = 1$.

Parameters of the data-generating process. In generating datasets, we varied seven parameters of the data-generating process:

1. Three levels for the total number of observations: $N = 200$, $N = 500$, $N = 1000$.

2. Two levels for the number of potential partitioning covariates U_1 through U_K: $K = 5$, $K = 15$ (where only U_1, U_2 and U_5 are true partitioning variables).

3. Two levels of intercorrelations between the covariates U_1 through U_K: $\rho_{U_k,U_{k'}} = 0.0$, $\rho_{U_k,U_{k'}} = 0.3$.

4. Three levels for the number of clusters: $M = 5$, $M = 10$, $M = 25$.

5. Three levels for the population standard deviation of the normal distribution from which the cluster specific intercepts are drawn: $\sigma_b = 0$, $\sigma_b = 5$, $\sigma_b = 10$.

6. Three levels for the intercorrelations between b and one of the U_k variables: b

and U_k uncorrelated, b correlated with a true partitioning covariate (i.e., U_2, U_1, or U_5, introducing a correlation of about 0.42), b correlated with a non-partitioning covariate (i.e., U_3 or U_4, introducing a correlation of about 0.42).

7. Two different levels for β_{j1}, the unstandardized mean difference in treatment outcomes, in subgroups with differential treatment effects. The levels for mean differences in subgroups with differential treatment effect were $|\beta_{j1}| = 2.5$ (corresponding to a medium effect size, Cohen's $d = 0.5$; Cohen, 1992) and $|\beta_{j1}| = 5.0$ (corresponding to a large effect size; Cohen's $d = 1.0$).

For each cell, 50 datasets with treatment-subgroup interactions were generated, resulting in $50 \times 3 \times 2 \times 2 \times 3 \times 3 \times 3 \times 2 =$ 32,400 training datasets. For the datasets without treatment-subgroup interactions, the 6th parameter of the data-generating process had only two levels (b correlated with one of the U_k variables, and b not correlated with any of the U_k variables). Therefore, $50 \times 3 \times 2 \times 2 \times 3 \times 3 \times 2 \times 2 =$ 21,600 datasets without treatment-subgroup interactions were generated.

It should be noted that when $\sigma_b = 0$, the correlation between b and one of the U_k variables is 0, by definition. However, datasets were created for this condition, to allow for a full factorial design of the simulation study; in reality, b and U are uncorrelated in these instances.

Variable distributions. As in Dusseldorp and Van Mechelen (2014), all covariates U_1 through U_K were drawn from a multivariate normal distribution with means μ_{U1}, μ_{U2}, μ_{U4}, and μ_{U5} fixed at 10, 30, -40, and 70, respectively. The means for all other covariates (i.e., μ_{U3}, and μ_{U6} through μ_{U15}) were drawn from a discrete uniform distribution on the interval $[-70, 70]$. All covariates U_1 through U_{15} have the same standard deviation: $\sigma_{Uk} = 10$. Correlations between the U_k variables vary according to the third facet of the simulation design described above.

To generate the random error term ϵ, for every observation we drew a value from a normal distribution with $\mu_\epsilon = 0$ and $\sigma_\epsilon = 5$.

To generate the cluster-specific intercepts b_m, we partitioned the sample into equally-sized clusters, conditional on one of the variables U_1 through U_5, producing the correlations in the sixth facet of the simulation design. For each cluster we drew a single b_m from a normal distribution with mean 0 and the value of σ_b given by the fifth facet of the simulation design. When b was correlated with one of the potential partitioning variables, the correlated potential partitioning variable was randomly selected.

To generate node-specific fixed effects, we partitioned the sample according to the terminal nodes of the tree in Figure 4.3. In combination with the seventh facet of the simulation design, this determines the values of β_j. For every observation, we generated a binomial variable (with probability .5) as an indicator for treatment type.

Finally, the response variable was calculated as the sum of the (node-specific) fixed effects, random effects and the error term: $y_i = x_i^\top \beta_j + z_i^\top b_m + \epsilon_i$.

Evaluation of performance

Tree size and accuracy. For every dataset, accuracy and size of the GLM and GLMM tree was evaluated. We calculated the total number of nodes in every tree, and compared it with the true tree size. For datasets without treatment-subgroup interactions, this allowed us to assess tree accuracy in terms of Type I error: the probability that the dataset is erroneously partitioned. For datasets with treatment-subgroup interactions, this allowed us to assess the probability that the dataset is erroneously not partitioned, and the extent to which the algorithms may detect spurious subgroups.

For datasets with treatment-subgroup interactions, we assessed the accuracy of the trees created by GLM and GLMM tree. An accurately recovered tree was defined as a tree with (1) the true tree size (i.e., total number of nodes equals 7), (2) the first split in the tree involving variable U_2 and a value of 30 ± 5, (3) the next split on the left involving variable U_1 and a value of 17 ± 5, and (4) the next split on the right involving variable U_5 and a value of 63 ± 5. Note that the allowance of ± 5 equals an allowance of plus or minus half the population standard deviation of the partitioning variable (σ_{U_k}).

To assess the impact of data-generating parameters on tree size for both algorithms, we performed ANOVAs with algorithm type and the parameters of the data-generating process as independent variables. In addition, interactions between algorithm type and each of the data-generating parameters were also entered as independent variables. The impact of predictors with main and/or interaction effects which explained a proportion of $> .01$ of variance were further investigated using graphical displays.

To assess the impact of data-generating parameters on tree accuracy in datasets with treatment-subgroup interactions, we used a GLM with algorithm type and the parameters of the data-generating process as independent variables. In addition, interactions between algorithm type and each of the data-generating parameters were also entered as independent variables. The impact of predictors with main and/or interaction effects with unstandardized regression coefficients $> .5$ (i.e., an in- or decrease in the log-odds of .5) were further investigated using graphical displays.

Predictive accuracy. We evaluated predictive accuracy of GLM and GLMM trees by calculating correlations between true and predicted treatment-effect differences (β_{j1} in Figure 4) for test observations. Note that this correlation was only assessed for datasets with treatment-subgroup interactions, as the true treatment differences have a constant value in datasets without treatment-subgroup interactions.

Using the same data for training and evaluation of a model results in overly optimistic estimates of predictive accuracy (Hastie, Tibshirani, & Friedman, 2009). Therefore, GLM and GLMM trees were used for prediction of new observations from test datasets. Test datasets were generated from the same population as the training datasets. Because the cluster-specific intercepts b were randomly generated for training as well as test datasets, test observations were from 'new' clusters. As a result, a model without random effects was used for prediction with GLMM tree.

For every dataset, correlation coefficients for each algorithm were calculated, representing the linear association between the true and predicted treatment-effect differences. To assess the impact of data-generating parameters on predictive accuracy, we performed ANOVAs with algorithm type and the parameters of the data-generating process as independent variables. In addition, interactions between algorithm type and each of the data-generating parameters were also entered as independent variables. The impact of predictors with main and/or interaction effects which explained a proportion of $> .01$ of variance were further investigated using graphical displays.

Software

R (R Core Team, 2014) was used for generation and analysis of all datasets. The `partykit` package (version 1.0-2; Hothorn & Zeileis, 2015) was employed for estimating GLM trees using the `lmtree` function for normal linear regressions and `glmtree` would be available for GLM trees with other response distributions. For estimation of GLMMs the `lmer` (or `glmer`, respectively) from the `lme4` package (version 1.1-7; Bates, Maechler, & Bolker, 2014) was employed using restricted maximum likelihood (REML) estimation.

For the estimation of GLMM trees the former two packages were combined in a new package `glmertree` (version 0.1-0; Fokkema & Zeileis, 2015; available from R-Forge). This provides functions `lmertree` and `glmertree` that iterate between estimation of the `lmtree`/`glmtree` model and the `lmer`/`glmer` model.

In all applications, the significance level α for the parameter instability tests in the trees was set to .05 with a Bonferroni correction applied for multiple testing. The minimum number of observations per node in the tree was set to 20 and the maximum tree depth was set to four, thus limiting the number of potential subgroups to eight. The iteration in the GLMM tree algorithm was repeated until the log-likelihood changes between two subsequent iterations fell below .001.

Results

Tree size in datasets without treatment-subgroup interactions. In Table 1, tree sizes for GLM and GLMM trees for datasets without treatment-subgroup interactions are presented. Overall, smaller trees were created by GLMM tree: the average tree size was 1.09 (SD = 0.44) for GLMM tree, and 2.02 (SD = 1.68) for GLM tree. The estimated probability that

Table 1: Tree size distributions for GLM and GLMM tree for datasets without treatment-subgroup interactions.

	tree size						
	1	3	5	7	9	11	total
GLMM tree	20625	932	43	0	0	0	21,600
	(.96)	(.04)	(< .01)	(.00)	(.00)	(.00)	(1.00)
GLM tree	14501	4202	2013	802	79	3	21,600
	(.67)	(.19)	(.09)	(.04)	(< .01)	(< .01)	(1.00)

Note. Bracketed values are proportions. Tree sizes are expressed as the total number of nodes in a tree. A tree with a total of J nodes has $(J + 1)/2$ terminal nodes; the true tree size in datasets without treatment-subgroup interactions was 1.

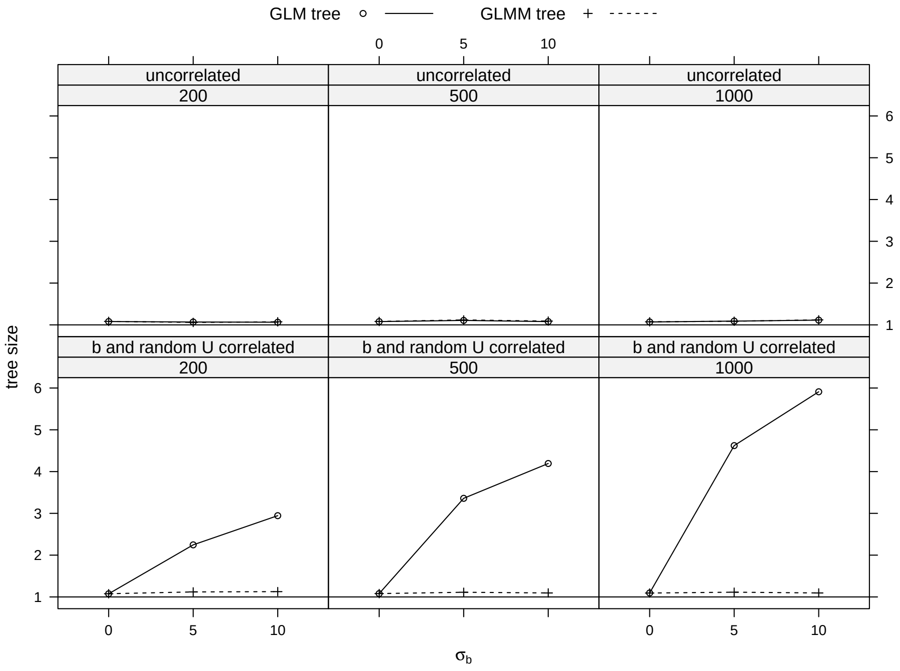

Figure 5. Average tree size of GLM and GLMM trees for datasets without treatment-subgroup interactions. Values 200, 500 and 1000 refer to sample size. Reference line at $y = 1$ represents the true tree size.

a dataset was erroneously partitioned was very small for GLMM tree (.04; Table 1), and much larger for GLM tree (.33; Table 1).

A graphical display was used to asses the effects of sample size, σ_b and the correlation

Table 2: Tree size distributions for GLM and GLMM tree for datasets with treatment-subgroup interactions.

	tree size							
	3	5	7	9	11	13	15	total
GLMM tree	2	281	29358	2675	83	1	0	32,400
	$(< .01)$	$(< .01)$	(.91)	(.08)	$(< .01)$	$(< .01)$	(.00)	(1.00)
GLM tree	139	955	20530	4565	4164	1619	428	32,400
	$(< .01)$	(.03)	(.63)	(.14)	(.13)	(.05)	(.01)	(1.00)

Note. Bracketed values are proportions. Tree sizes are expressed as the total number of nodes in the tree. A tree with a total of J nodes has $(J + 1)/2$ terminal nodes; the true tree size in datasets with treatment-subgroup interactions was 7.

between b and one of the U_k variables, on tree size (Figure 5). When random effects were absent (i.e., $\sigma_b = 0$), both GLM and GLMM tree tend to create trees of size 1. In the presence of random effects, GLMM tree also tends to create trees of size 1, but GLM tree created much larger trees, when b was correlated to one of the U_k variables. This effect was stronger when sample size was larger.

Tree size in datasets with treatment-subgroup interactions. In datasets with treatment-subgroup interactions, GLMM trees were also smaller than GLM trees. For these datasets, the true tree size was 7 (4 terminal nodes and 3 inner nodes; Figure 4). Th distribution of tree sizes for GLM and GLMM tree in datasets with treatment-subgroup interactions are presented in Table 2. The average size of GLMM trees was 7.16 (SD = 0.62), and the average size of GLM trees was 8.12 (SD = 2.05). The estimated probability that a dataset was erroneously not partitioned was 0, for both GLM and GLMM tree. However, Table 2 shows that a proportion of .91 of GLMM trees matched the true tree size, whereas a proportion of only .63 of GLM trees matched the true tree size (Table 2).

A graphical display was used to assess the effects of sample size, σ_b and the correlation between b and one of the U_k variables, on tree size (Figure 6). When random effects were absent (i.e., $\sigma_b = 0$), both GLM and GLMM tree created trees with a size of about 7, on average.

Clear differences in performance between GLM and GLMMtree were observed when $\sigma_b > 0$. When b is not correlated with one of the U_k variables, when sample size is small (i.e., 200) and when σ_b is large (i.e., 10), GLM tree has difficulty detecting splits and grows trees that are too small, on average. When b is not correlated with one of the U_k variables and when sample size is larger (i.e., 500 or 1000), GLM and GLMM trees are about the same size (i.e., $\approx$ 7). When b is correlated with one of the U_k variables, GLM creates spurious splits, especially when sample size is larger (i.e., 500 or 1000) and when σ_b is large

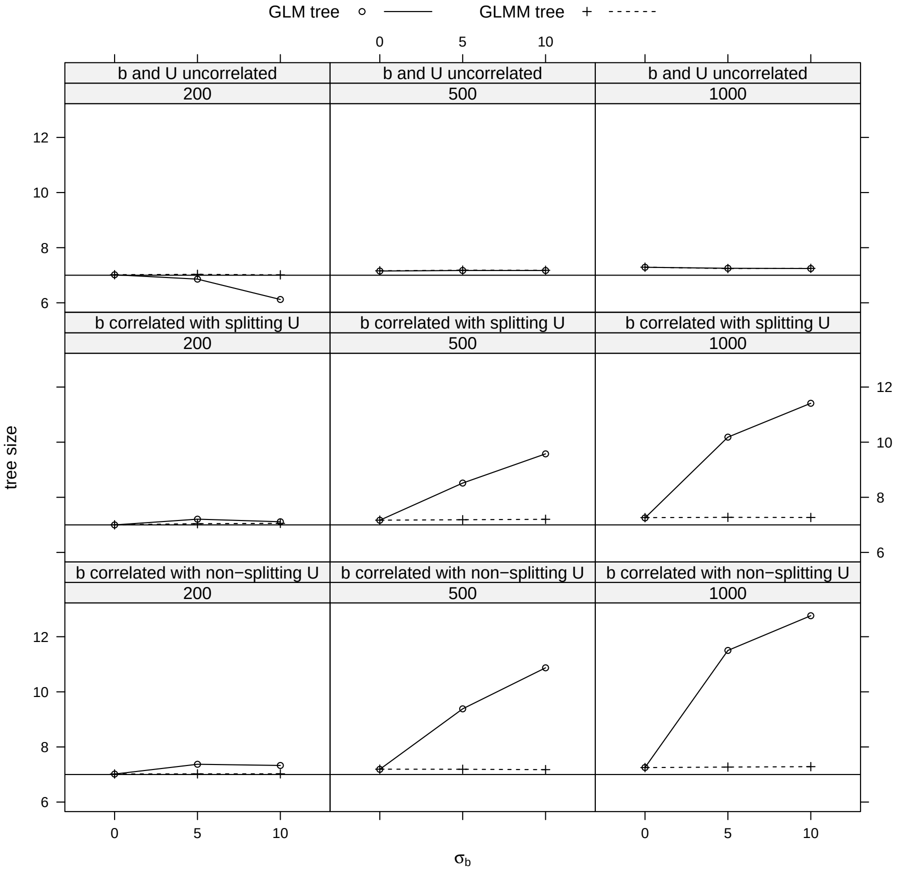

Figure 6. Average tree size of GLM and GLMM trees for datasets with treatment-subgroup interactions. Values 200, 500 and 1000 refer to sample size. Reference line at $y = 7$ represents true tree size.

(i.e., 10). This effect was stronger when b was correlated to a non-splitting variable.

Tree accuracy in datasets with treatment-subgroup interactions. To assess the accuracy of the trees created by GLM and GLMM tree, we inspected the variables and values that were selected for partitioning in every dataset. For the first split, GLMM tree always selected the true partitioning variable (U_2). GLM tree selected a wrong partitioning variable (I.e., U_1 in only one dataset. The true splitting value for U_2 was 30 (Figure 4), and the mean splitting value selected for the first split (involving U_2) was 29.94, for both GLM and GLMM tree. However, GLM tree showed somewhat higher variability in recovering the splitting value for the first split (involving U_2), than did GLMM tree (SD = 0.154 and SD = 0.127, respectively).

Overall, GLMM tree performed well in recovering treatment-subgroup interactions,

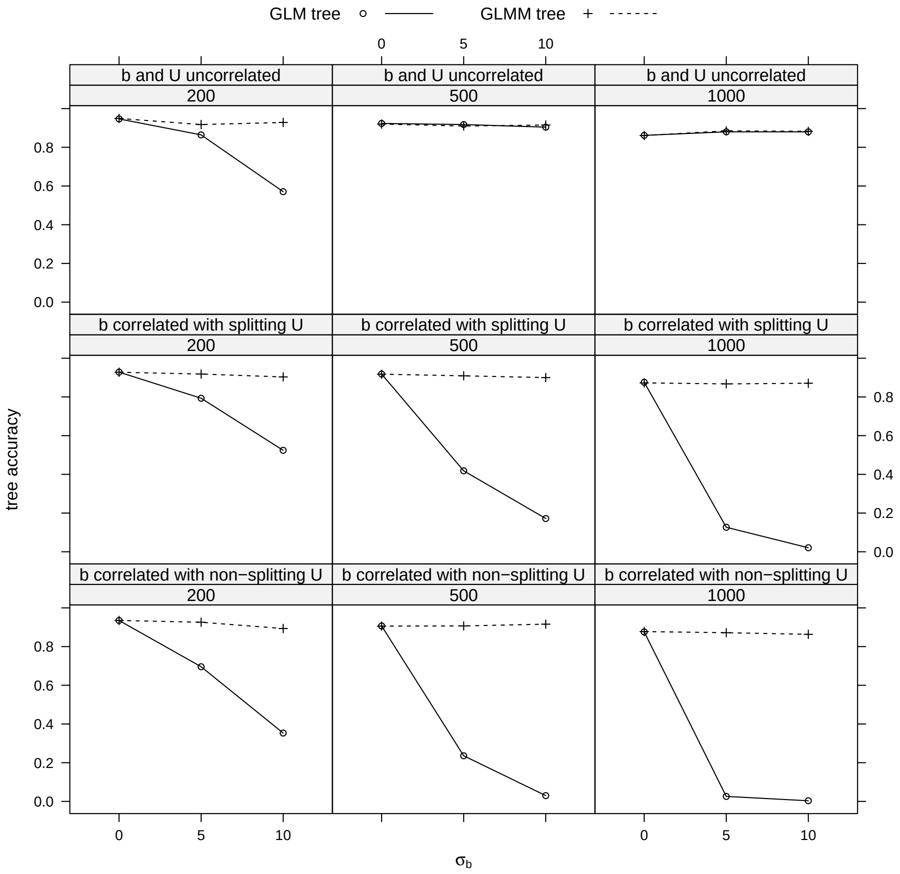

Figure 7. Average accuracy of GLM and GLMM trees. Accuracy of trees is defined as the proportion of datasets in which the true tree was accurately recovered. Values are correlated to one of the U_k variables; values 200, 500 and 1000 refer to sample size.

accurately recovering the tree in 90.19% of datasets. GLM tree accurately recovered the treatment-subgroup interactions in only 61.44% of datasets.

A graphical display was used to assess the effects of sample size, σ_b and the correlation between b and one of the U_k variables, on the probability of accurate tree recovery for both algorithms (Figure 7). When random effects were absent from the datasets (i.e., $\sigma_b = 0$), the trees recovered by GLM and GLMM tree were equally accurate, on average. In the presence of random effects, GLM trees were much less accurate than GLMM trees. This was found for all sample sizes, when b was correlated to one of the U_k variables, and the effect was somewhat stronger when the correlated U_k was not a true partitioning variable. When b was not correlated to one of the U_k variables, GLMM tree clearly outperformed GLM tree only when sample size was small (i.e., 200).

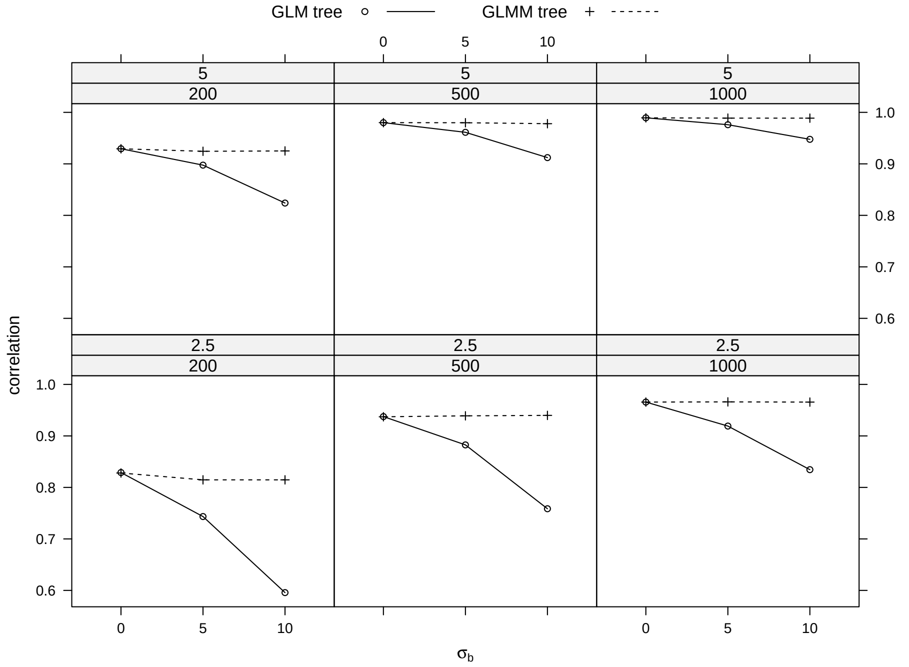

Figure 8. Average predictive accuracy of GLM and GLMM trees. Predictive accuracy of trees is defined as the correlation between the true and predicted differences between Treatment 1 and 2. Values 5 and 2.5 refer to the absolute value of the unstandardized treatment-effect difference in subgroups with treatment-effect differences; values 200, 500 and 1000 refer to sample size.

Predictive accuracy on test data

To assess predictive accuracy of both algorithms, correlation between the true and predicted treatment-effect differences of both algorithms were calculated for every dataset. Overall, treatment-effect differences predicted by GLMM tree were closer to the true differences than those predicted by GLM tree. The average correlation between the true and predicted treatment-effect differences over all 32,400 datasets was .88 (SD = 0.19) for GLM tree, and .94 (SD = 0.11) for GLMM tree.

A graphical display was used to assess the effects of sample size, σ_b and the correlation between b and one of the U_k variables, on the predictive accuracy of both algorithms (Figure 8). Both algorithms showed higher predictive accuracy when sample size was larger, and when treatment-effect differences were larger. When random effects were absent from the datasets (i.e., $\sigma_b = 0$), predictions of GLM and GLMM tree were equally accurate. In the presence of random effects, GLM tree predictions were always much less accurate than those of GLMM tree. This effect was stronger when σ_b was larger, sample size was larger, and/or treatment-effect differences were larger.

Application to patient-level meta-analysis dataset on the effects of treatments for depression

Method

To illustrate the application of, and differences in the results of GLM tree and GLMM tree, we apply both algorithms to a dataset from a meta-analytic study of Cuijpers et al. (2014). This meta-analysis was based on individual-patient data from 14 RCTs, comparing the effects of psychotherapy (cognitive behavioral therapy; CBT) and pharmacotherapy (PHA) in the treatment of depression. The study of Cuijpers et al. (2014) was aimed at establishing whether gender is a predictor or moderator of the outcomes of psychological and pharmacological treatments for depression. Treatment outcomes were assessed by means of the 17-item Hamilton Rating Scale for Depression (HAM-D; Hamilton, 1960). Cuijpers et al. (2014) found no indication that gender either predicted or moderated treatment outcomes. Further details on the dataset are provided in Cuijpers et al. (2014).

In our analyses, posttreatment HAM-D score is the outcome variable, and potential partitioning variables are age, gender, level of education, presence of a comorbid anxiety disorder at baseline, and pretreatment HAM-D score. The predictor variable in the linear model was treatment type (0 = CBT and 1 = PHA). An indicator for study was used as the cluster indicator.

In RCTs, treatment effects are often estimated after controlling posttreatment values on the outcome measure for the linear effect of pretreatment values on the same measure. Therefore, we included the predictions of a linear regression of HAM-D posttreatment on HAM-D pretreatment scores, as an offset variable in all models. An offset variable is a linear predictor with an a-priori determined coefficient of one. Including the linear regression predictions as an offset has the same effect as statistically controlling for the linear effects of pretreatment scores, as is often done in ANCOVA.

We build all trees using data of patients with complete observations; that is, observations with non-missing values for potential partitioning variables, and pre- and posttreatment HAM-D score. As a result, data from 694 patients from 7 studies were included for the analyses. Results of our analysis may therefore not be representative of the complete dataset of the meta-analysis by Cuijpers et al. (2014).

The predictive accuracy of GLM and GLMM tree was assessed by calculating the average correlations between observed and predicted HAM-D scores, based on 50-fold cross validation.

Results

The trees resulting from application of GLM and GLMM tree to the dataset are presented in Figure 9 and 10, respectively. Note that the GLM tree in Figure 9 is also the

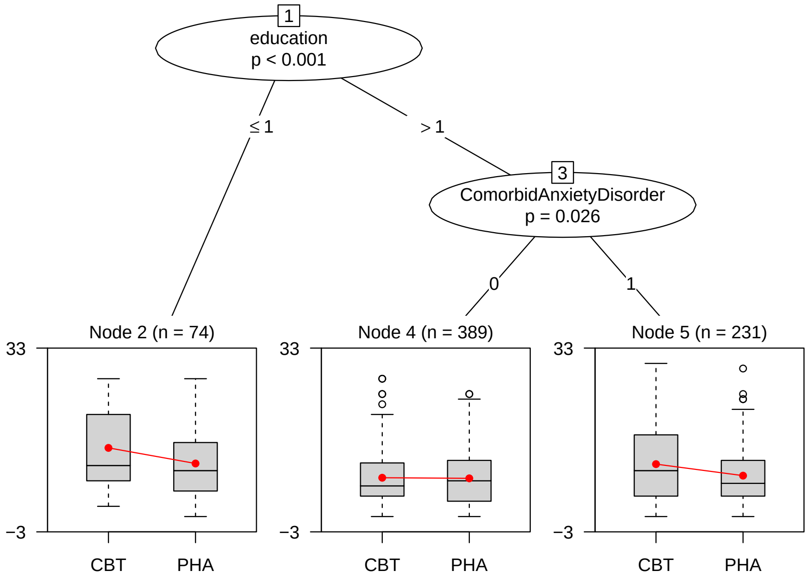

Figure 9. GLM tree for prediction of posttreatment total scores on the Hamilton Rating Scale for Depression (HAM-D). The y-axes of the boxplots represent posttreatment HAM-D scores, and the x-axes represent treatment levels: cognitive behavior therapy (CBT) vs. pharmacotherapy (PHA).

tree that is created in the first iteration of the GLMM tree algorithm.

The GLM tree (Figure 9) selected level of education as the first partitioning variable, and presence of a comorbid anxiety disorder as a second partitioning variable, for observations with a higher level of education. Node 2 of Figure 9 indicates that for patients with a low level of education, antidepressant medication provides the greatest reduction in HAM-D scores. Node 4 indicates that for patients with a higher level of education, and no comorbid anxiety disorder, the reduction in HAM-D scores is about the same for CBT and antidepressant mediation. Node 5 indicates, that for patients with a higher level of education and a comorbid anxiety disorder, the reduction in HAM-D scores is greatest for pharmacotherapy.

By taking into account the study-specific intercepts, the final GLMM tree (Figure 10) suggests that the first split made by GLM tree is a spurious split. The GLMM tree selected only presence of a comorbid anxiety disorder as a partitioning variable. The terminal nodes of Figure 10 show only a single treatment-subgroup interaction: for patients without a comorbid anxiety disorder, CBT and antidepressant medication provide more or less the same reduction in HAM-D scores, whereas for patients with a comorbid anxiety disorder, antidepressant medication provides a greater reduction in HAM-D scores. The estimated intraclass correlation coefficient for the random intercepts was .05.

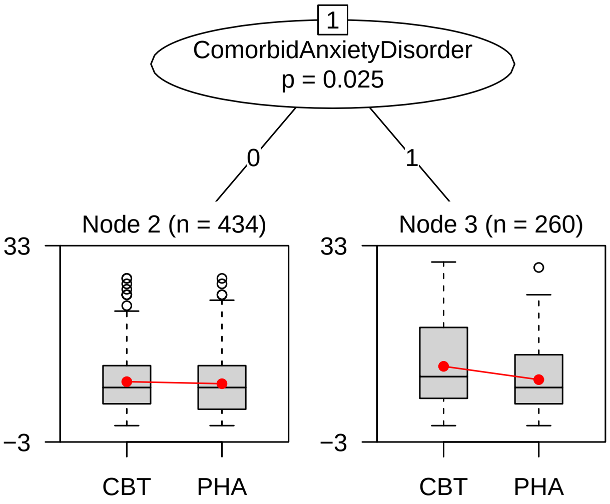

Figure 10. GLMM tree for prediction of posttreatment total scores on the Hamilton Rating Scale for Depression (HAM-D). The y-axes of the boxplots represent posttreatment HAM-D scores, and the x-axes represent treatment levels: cognitive behavior therapy (CBT) vs. pharmacotherapy (PHA).

Assessment of predictive accuracy by means of 50-fold cross validation showed that the GLMM tree had higher predictive accuracy than the GLM tree. The correlation between true and predicted posttreatment HAM-D total scores, averaged over the 50 folds, was .28 (var = .067) for GLMM tree, and .19 (var = .084) for GLM tree. This indicates that GLMM tree provided higher predictive accuracy, on average, and also somewhat lower variability of predictive accuracy than GLM tree.

Discussion

In the current paper, we presented the GLMM tree algorithm, which allows for the estimation of a GLM-based recursive partition, as well as the estimation of random-effects parameters. The results of our simulation study show that GLMM tree performed very well in recovering treatment-subgroup interactions, by recovering the true tree structure in 90% of the simulated datasets with treatment-subgroup interactions. In contrast, GLM tree recovered the true tree structure in only 61% of the datasets with treatment-subgroup interactions. In the absence of treatment-subgroup interactions, GLMM tree erroneously detected subgroups in only 4% of the datasets, whereas GLM tree erroneously detected subgroups in 33% of those datasets. In other words, the Type I error rate of GLMM tree very closely resembled the α level used for evaluating significance of parameter instability, whereas the Type I error rate of GLM tree clearly exceeded this value.

The better performance of GLMM tree was mostly observed when random effects in the datasets were sizable, and random intercepts were correlated with potential partitioning variables. In these instances, the random effects gave rise to spurious subgroup detection (spurious splits) by GLM tree, both in datasets with and without treatment-subgroup interactions.

Also, predictive accuracy of GLMM tree was higher than that of GLM tree. The average correlation between the true treatment differences and those predicted by GLMM tree was .94. The average correlation between the true treatment differences and those predicted by GLM tree was .88. In terms of predictive accuracy, GLMM tree clearly outperformed GLM tree when random effects in the datasets were sizable, and the differences in treatment effects were relatively small (i.e., $d = .5$).

As expected, when random effects were absent from the simulated datasets, GLM tree and GLMM tree showed high and equal predictive accuracy. This finding indicates that GLMM tree can be applied, whenever cluster-specific random effects are expected. In the absence of random effects, GLM tree and GLMM tree are expected to perform equally well, and in the presence of random effects, GLMM tree will outperform GLM tree. This is especially the case with large sample sizes ($N > 200$), as the increased power will likely cause GLM tree to create spurious splits in the presence of random effects.

Not surprisingly, for both algorithms, accuracy of predicted treatment differences was less when sample size was low (i.e., $N = 200$). Sample size influenced performance of GLM tree and GLMM tree similarly, suggesting that a larger number of estimated parameters for GLMM tree does not adversely influences accuracy at low sample sizes. Our simulation results do warrant some caution for the detection of treatment-subgroup interactions or treatment moderators in small datasets (e.g., single RCTs), but irrespective of the algorithm used.

These findings are encouraging for the use of GLMM tree in the detection of treatment-subgroup interactions in datasets with clustered structures. However, it should be noted that the simulations show that GLMM tree performs very well, if the model is correctly specified. That is, if there are subgroups with respect to the partitioning variables, so that there are different parameters of the GLM in each of these subgroups, then the algorithm will accurately recover those subgroups. However, misspecification of the model can reduce performance. One source of misspecification would be, when relevant variables are not included in the GLM or as partitioning variables. If there are actual subgroups, but the variables describing them are not entered as partitioning variables, the algorithm can only approximate the subgroups using the partitioning variables that are available. Or, if the coefficients of other variables vary across subgroups, then those variables should also be included in the linear predictor of the GLM. Another source of misspecification would be the inclusion of irrelevant variables, either in the linear predictor of the GLM or as partitioning variables, which may reduce the power to detect the actual subgroups. However, it should be noted that in our simulations, the number of partitioning variables did not substantially influence performance of the algorithm(s).

In conclusion, GLMM tree provided highly accurate recovery of treatment-subgroup interactions and predictions of treatment effect differences, both in the presence and absence of cluster-specific random effects. Therefore, GLMM tree is a promising algorithm for the detection of treatment-subgroup interactions in datasets with a clustered structure, like for example in multi-center trials, individual-level patient data meta-analyses, and longitudinal studies.

References

Bates, D., Maechler, M., & Bolker, B. (2014). *lme4: Linear mixed-effects models using S4 classes.* Retrieved from `http://CRAN.R-project.org/package=lme4` (R package version 1.1-7)

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees.* New York: Wadsworth.

Bryk, A. S., & Raudenbush, S. W. (1992). *Hierarchical linear models: Applications and data analysis methods.* Newbury Park, CA: Sage.

Cohen, J. (1992). A power primer. *Psychological Bulletin*, *112*(1), 155–159.

Cooper, H., & Patall, E. A. (2009). The relative benefits of meta-analysis conducted with individual participant data versus aggregated data. *Psychological Methods*, *14*(2), 165.

Cuijpers, P., Weitz, E., Twisk, J., Kuehner, C., Cristea, I., David, D., ... Hollon, S. D. (2014). Gender as predictor and moderator of outcome in cognitive behavior therapy and pharmacotherapy for adult depression: An "individual-patients data" meta-analysis. *Depression and Anxiety*, *31*(11), 941–951.

Doove, L. L., Dusseldorp, E., Van Deun, K., & Van Mechelen, I. (2014). A comparison of five recursive partitioning methods to find person subgroups involved in meaningful treatment-subgroup interactions. *Advances in Data Analysis and Classification*, *8*, 403–425.

Driessen, E., Smits, N., Peen, J., Don, F. J., Kool, S., Westra, D., ... Van, H. L. (2014). *Differential efficacy of cognitive behavioral therapy and psychodynamic therapy for major depression: A study of prescriptive factors.* Manuscript under review.

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

Dusseldorp, E., & Van Mechelen, I. (2014). Qualitative interaction trees: A tool to identify qualitative treatment-subgroup interactions. *Statistics in Medicine*, *33*(2), 219–237.

Fokkema, M., & Zeileis, A. (2015). *glmertree: Generalized linear mixed model trees.* Retrieved from `http://R-Forge.R-project.org/R/?group`$_id = 261$ (R package version 0.1-0)

Foster, J. C., Taylor, J. M. G., & Ruberg, S. J. (2011). Subgroup identification from randomized clinical trial data. *Statistics in Medicine*, *30*(24), 2867–2880.

Hajjem, A., Bellavance, F., & Larocque, D. (2011). Mixed effects regression trees for clustered data. *Statistics & Probability Letters*, *81*(4), 451–459.

Hamilton, M. (1960). A rating scale for depression. *Journal of Neurology, Neurosurgery and Psychiatry*, *23*(1), 56.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning* (2nd ed.). New York: Springer.

Higgins, J., Whitehead, A., Turner, R. M., Omar, R. Z., & Thompson, S. G. (2001). Meta-analysis of continuous outcome data from individual patients. *Statistics in Medicine*, *20*(15), 2219–2241.

Hothorn, T., & Zeileis, A. (2015). partykit: A modular toolkit for recursive partytioning in R. *Journal of Machine Learning Research*. (Forthcoming, preprint at http://EconPapers.RePEc.org/RePEc:inn:wpaper:2014-10)

Koopman, L., Van der Heijden, G. J. M. G., Glasziou, P. P., Grobbee, D. E., & Rovers, M. M. (2007). A systematic review of analytical methods used to study subgroups in (individual patient data) meta-analyses. *Journal of Clinical Epidemiology*, *60*(10).

Kraemer, H. C., Frank, E., & Kupfer, D. J. (2006). Moderators of treatment outcomes: Clinical, research, and policy importance. *Journal of the American Medical Association*, *296*(10), 1286–1289.

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

Merkle, E. C., & Zeileis, A. (2013). Tests of measurement invariance without subgroups: A generalization of classical methods. *Psychometrika*, *78*(1), 59–82.

R Core Team. (2014). R: A language and environment for statistical computing [Computer software manual]. Vienna, Austria. Retrieved from `http://www.R-project.org/`

Sela, R. J., & Simonoff, J. S. (2012). RE-EM trees: A data mining approach for longitudinal and clustered data. *Machine Learning*, *86*(2), 169–207.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application, and characteristics of classification and regression trees, bagging, and random forests. *Psychological Methods*, *14*(4), 323.

Su, X., Tsai, C.-L., Wang, H., Nickerson, D. M., & Li, B. (2009). Subgroup analysis via recursive partitioning. *The Journal of Machine Learning Research*, *10*, 141–158.

Van den Noortgate, W., Opdenakker, M. C., & Onghena, P. (2005). The effects of ignoring a level in multilevel analysis. *School Effectiveness and School Improvement*, *16*(3), 281-203.

Zeileis, A. (2005). A unified approach to structural change tests based on ML scores, F statistics, and OLS residuals. *Econometric Reviews*, *24*(4), 445–466.

Zeileis, A., & Hornik, K. (2007). Generalized M-fluctuation tests for parameter instability. *Statistica Neerlandica*, *61*(4), 488–508.

Zeileis, A., Hothorn, T., & Hornik, K. (2008). Model-based recursive partitioning. *Journal of Computational and Graphical Statistics*, *17*(2), 492–514.

Appendix: Notation

$1, \ldots, i, \ldots, N$	observation number
$1, \ldots, j, \ldots, J$	terminal node number in a tree
$1, \ldots, k, \ldots, K$	partitioning variable number
$1, \ldots, m, \ldots, M$	cluster number
β_j	column vector of fixed-effects coefficients in terminal node j
b_m	column vector of random-effects coefficients in cluster m
d_j	$\beta_{j1}/\sigma_\epsilon$; effect size of treatment-effect differences between Treatment 1 and Treatment 2 in terminal node j
ϵ	deviation of observed treatment outcome y from its expected value
r	iteration number
σ_b	standard deviation of b
σ_ϵ	standard deviation of ϵ
U_k	(potential) partitioning variable k
x_i	column vector of fixed-effects predictor variable values for observation i
y_i	treatment outcome for observation i
z_i	column vector of random-effects predictor variable values for observation i