MetaForest: Exploring heterogeneity in meta-analysis using random forests

Caspar J. van Lissa

Erasmus University Rotterdam

Author Note

Abstract

Meta-analyses in psychology often lack the power to adequately account for between-studies heterogeneity. The number of studies on any topic is typically low, because research is cost- and time-intensive. At the same time, a host of potential moderators are introduced when similar research questions are examined in different labs, sampling from different populations, using idiosyncratic methods and instrumentation. Such between-studies heterogeneity presents a substantial challenge to data aggregation in classic meta-analysis. When the causes for heterogeneity are known a-priori, they can be accounted for using meta-regression. What is currently lacking, however, is an exploratory approach, to be used when heterogeneity is suspected, but it is not known which moderators most strongly influence the observed effect size. Recently, weighted regression trees have been used to explore heterogeneity in meta-analysis. Although this provides a promising first step, single trees have many limitations, which can be overcome by using random forests: A powerful learning algorithm, which is flexible, yet relatively robust to overfitting. The present paper introduces MetaForest: An adaptation of random forests for meta-analysis. We present two simulation studies, which illustrate that, in datasets as small as 20 cases, MetaForest outperforms single trees, in terms of three metrics: 1) Predictive performance; 2) power, as evidenced by the proportion of datasets in which the algorithm achieved a positive R2cv; and 3) the ability to distinguish relevant moderators from irrelevant moderators, using variable importance measures. We discuss how MetaForest can enhance the exploration of between-studies heterogeneity when conducting meta-analyses in diverse bodies of literature.

Keywords: meta-analysis, random forests, CART, MetaForest, metaCART

MetaForest: Exploring heterogeneity in meta-analysis using random forests

In recent years, much attention has been devoted to the perceived "replicability crisis" in psychology: The crisis of confidence over the reliability of published results in our field (Maxwell, Lau, & Howard, 2015; Simmons, Nelson, & Simonsohn, 2011). Several remedies have been proposed to derive knowledge that will stand the test of time. One proposed solution to establish the reliability of published findings is meta-analysis: Aggregating the findings of multiple studies by computing a summary effect size across studies (Braver, Thoemmes, & Rosenthal, 2014; Laws, 2016). A requirement of meta-analysis is that the studies being aggregated are conceptually similar, and ideally, close replications (Fabrigar & Wegener, 2016; Higgins, Thompson, & Spiegelhalter, 2009; Maxwell et al., 2015). However, in psychology and other fields, there is often substantial heterogeneity between studies on the same topic. Similar research questions are studied in different laboratories, using different methods, instruments, and samples. These differences between studies are known as "moderators", and methods have been developed to account for their influence. However, extant approaches lack the power to assess more than a handful of known moderators, or to investigate interactions between moderators, and non-linear effects. Heterogeneity between studies thus presents a non-trivial challenge to data aggregation using classic meta-analytic methods. At the same time, it also offers an unexploited opportunity to learn which differences between studies have an impact on the effect size found, if adequate exploratory techniques can be developed. Recently, a first step towards this end has been proposed: Meta-CART, a technique which uses a classification/regression tree to group studies into clusters based on combinations of study characteristics that jointly predict effect size, and then conducts subgroup meta-analysis, using cluster membership as a moderator. Although this approach is promising, single trees have several limitations. These can be overcome by using random forests: A technique that grows many single trees, and aggregates their predictions. The present paper introduces MetaForest, a new technique for meta-analysis on heterogeneous bodies of literature. Using simulation studies, it is

demonstrated that MetaForest overcomes the limitations of meta-CART, and has substantial power to detect heterogeneity between studies.

Classic meta-analytic approaches

There are two classic approaches to meta-analysis. One is fixed-effects meta-analysis, which assumes that each observed effect size is an estimate of an underlying true effect size, and is subject to sampling error (Hedges & Vevea, 1998). Thus, for a collection of k studies, the observed effect size y_i of each individual study i (for $i = 1, 2, \ldots k$) is given by:

$$y_i = \theta_i + \epsilon_i \quad \text{where } \epsilon_i \sim N(0, \sigma_i^2) \tag{1}$$

The second approach is random-effects meta-analysis, which assumes that the true effect size follows a distribution, and consequently, that differences between observed effect sizes arise from two sources of variance: Sampling error, and the deviation of each study from the mean of a distribution of true effect sizes, or between-studies variance (Borenstein, Hedges, Higgins, & Rothstein, 2009; Hedges & Vevea, 1998). The random-effects model is thus given by:

$$\left.\begin{aligned} y_i &= \theta_i + \epsilon_i \quad \text{where } \epsilon_i \sim N(0, \sigma_i^2) \\ \theta_i &= \mu + \zeta_i \quad \text{where } \zeta_i \sim N(0, \tau^2) \end{aligned}\right\} \tag{2}$$

In both of these meta-analytic approaches, the summary effect is a weighted mean of the observed effect sizes. The weights assigned to individual studies are based on the sources of variance assumed to influence the observed effect size. Fixed-effects meta-analysis assumes that the only source of variance affecting the observed effect sizes is sampling error. This implies that studies with large sample sizes provide more accurate estimates of their underlying true effect, and should contribute more to the weighted mean than studies with a smaller sample. Thus, fixed-effect weights are defined as the reciprocal of the effect size variances:

$$W_i = \frac{1}{\sigma_i^2} \tag{3}$$

If the additional assumption can be made that each study in the sample is assessing the same underlying true effect (in other words, that $\theta_1 = \theta_2 \ldots \theta_k$), then the weighted

average of a fixed-effects meta-analysis can be considered a point estimate of the true effect size. If this assumption is not made, the weighted average should be considered a summary of the observed effect sizes (Hedges & Vevea, 1998).

Random-effects meta-analysis instead assumes that the underlying true effect sizes follow a distribution. Each study — even smaller ones — thus provides some information about this distribution. Mathematically, this is implemented by attenuating study weights relative to the estimated amount of between-study variance. If the estimated amount of between-study variance is negligible, the weights are thus identical to the fixed-effect weights. As the between-study variance increases, the weights are increasingly attenuated, until they approach unity. In other words, the more heterogeneous observed effect sizes are, the more equally each study contributes to the weighted mean:

$$W_i^* = \frac{1}{\sigma_i^2 + \hat{\tau}^2} \tag{4}$$

The summary effect in random-effects analysis is a point estimate of the true effect size distribution's mean.

It has been argued that the fixed effect model rarely holds in the social sciences. Human behavior is notoriously complex (Earp & Trafimow, 2015), and consequently, any psychological phenomenon is likely subject to a host of potential moderators (Cesario, 2014). Studies examining the identical, or similar, research questions often differ in terms of research populations, methodology, and instrumentation (Higgins et al., 2009). Even replication studies, which are by definition designed to be equivalent, typically display heterogeneity in effect sizes due to unforeseen moderators (Maxwell et al., 2015; Simmons et al., 2011). Three approaches have been proposed to deal with heterogeneity (Higgins et al., 2009): First, if studies are assumed to be different and unrelated, they should not be meta-analyzed. Secondly, if studies can be assumed to be similar, a random-effects model can be used to estimate the distribution of the true effect size. Thirdly, if the moderators responsible for between-studies heterogeneity are known a priori, their influences can be modeled using meta-regression. To this end, the models described in Equations 1 and 2 are extended by specifying a linear model for the

parameters θ_i, or μ, respectively. Thus, to account for heterogeneity introduced by moderator x_1 using random-effects meta-regression, one would specify that $\mu_i = \beta_0 + \beta_1 x_{1i}$. In meta-regression under the random-effects model, the error term ζ_i captures residual heterogeneity after accounting for the moderators (Higgins et al., 2009). What is lacking in the literature, however, is a feasible approach for situations where between-studies heterogeneity is suspected, and moderators have been measured, but the specific causes of heterogeneity are not known a priori. The present study aimed to address this problem.

Software to conduct meta-analysis with multiple moderators is freely available, and could readily be used to explain why different studies obtain different effects (metafor; Viechtbauer, 2010). Nevertheless, published meta-analyses rarely account for more than a few moderators, and some do not include moderators at all. One possible explanation for this is, that the number of studies in meta-analyses is often too low to obtain the power required to examine heterogeneity reliably (Riley, Higgins, & Deeks, 2011). Moreover, there is a paucity of theory regarding sources of heterogeneity to help whittle the long list of potential moderators down to a manageable number (Thompson & Higgins, 2002). Many meta-analysts are thus faced with what is known as the "curse of dimensionality": The problem that arises when the number of variables to be considered is large, relative to the number of cases in the data. In meta-analysis, the number of studies available is often low, because conducting research is cost- and time-intensive, and yet there are many potential sources of heterogeneity, resulting in a high number of moderators to be considered. Such cases do not fit comfortably into the classic meta-analysis paradigm, which, like any regression-based approach, requires a high number of cases per parameter. Instead, this problem calls for an exploratory technique which can perform variable selection — indentifying which moderators most strongly influence the observed effect size.

Recently, a method has been proposed to identify interactions between moderators in meta-analysis. This technique, called "metaCART", uses regression/classification trees to identify which combinations of behavior change techniques jointly predict the

effectiveness of treatment in meta-analyses of intervention studies. The classification/regression tree algorithm works by splitting the dataset repeatedly into ever more homogenous groups with regard to the outcome (effect size). Starting with the full sample, it finds the moderator and splitting value which maximize the homogeneity of the effect size in the two groups resulting from the split. The resulting groups are split again recursively, until a pre-specified stopping criterion is reached: Most commonly, when all of the resulting groups are perfectly homogenous, or when one of the groups contains a pre-specified minimum number of observations (for an accessible introduction, see Strobl, Malley, & Tutz, 2009). The key advantage of tree models is that they have higher power than linear regression in situations where moderators outnumber observations. Moreover, tree models perform variable selection: Any variable unrelated to the outcome will not be chosen for a split. Another notable advantage of trees is their ability to handle interactions between moderators, and non-linear effects. Tree models thus have greater flexibility in modeling the complexity of human behavior than linear models do (Earp & Trafimow, 2015). Real-data examples and simulation studies suggest that metaCART might be useful for detecting interactions between moderators in meta-analysis (Dusseldorp, van Genugten, van Buuren, Verheijden, & van Empelen, 2014; Li, Dusseldorp, & Meulman, 2017). A potentially more important application of tree-based approaches to meta-analysis, however, relates to the larger problem of dealing with heterogeneity in meta-analysis.

Adapting single trees for meta-analysis is relatively straightforward: Just as with classic meta-analysis, effect sizes can be weighted according to the reciprocal of their variance (using either the fixed- or random-effects weights). However, whereas classic meta-analysis calculate a weighted average effect size, tree models implement these weights in the impurity function that is used to evaluate the homogeneity of the groups resulting from a potential split. Studies with larger weights thus exert more influence in determining what splits are made. MetaCART (Li et al., 2017) follows a two-step approach: First, weighted classification/regression trees are used to group studies into clusters with similar effect sizes. The resulting tree model can be visualized, and is

interpreted as a visual representation of how moderators interact to predict effect size (Dusseldorp et al., 2014; Li et al., 2017). In the second step, "cluster membership", i.e., which of the tree's terminal nodes a specific study ends up in, is used as a categorical moderator in a mixed-effects subgroup meta-analysis. Mixed-effects subgroup analysis is a type of meta-regression, which assumes random effects within groups, and a fixed effect across groups. In other words, all differences in effect sizes between subgroups are assumed to be explained by the subgroup membership, and any unexplained variance is assumed to be sampling error. The subgroup analysis from the second step is then used to test the significance of differences between the final groups.

Limitations of tree models

Although the tree-based aproach constitutes a promising first step towards addressing the problem of heterogeneity in meta-analysis, several limitations remain. First, although the apparent interpretability of tree models has undoubtedly contributed to their popularity. However, this interpretability is, to some extent, misleading. Tree models suffer from an inherent instability, in the sense that minor variations in the data used to build the model can lead to major changes in tree structure (Hastie, Tibshirani, & Friedman, 2009). The effect of this instability is amplified, because trees are susceptible to order effects: Trees are hierarchical structures, so if a different variable or value is selected for one split in the tree, this influences all down-stream splits. Furthermore, because the algorithm can only makes binary splits, it has difficulty capturing linear, additive effects. These, too, will be represented as a sequence of "interactions". Finally, variables' recurrence in a single tree cannot be taken as evidence for variable importance: Imagine two correlated moderator variables, x_1 and x_2, both of which are associated with effect size. If x_1 is selected in one split, x_2 is less likely to be selected downstream. This might lead a researcher to conclude that x_2 is unimportant, although — in reality — both variables are equally important. These shortcomings imply that the true data-generating model may not be evident from the tree structure (see also Strobl et al., 2009). To summarize, the tree structure is readily interpretable in terms of how values of predictors are mapped to the

predicted outcome, but it does not necessarily reflect the true data-generating model, nor the importance of the variables used at each split.

Tree-based models are also very prone to overfitting (Hastie et al., 2009): The problem that arises when a model learns too much from the data used to build the model (training data), and ends up representing not only systematic patterns in the data, but also its idiosyncrasies (random variation). An overfit model generalizes poorly to new data (testing data), because its predictions are partly based on these idiosyncrasies. Overfitting is commonly controlled by pruning the tree, typically to a smaller tree that minimizes some form of cross-validation error. However, it is important to note that, while this improves predictions in future data, it does not improve the tree's ability to uncover the true model. Pruned trees remain equally susceptible to the aforementioned instability and order effects. One final caveat of using single trees is the fact that they provide piecewise constant predictions: Each terminal node of the tree predicts a specific constant value for the outcome. Predictions can thus jump substantially for slight changes in the values of moderators, if this causes a study to be assigned into a different group. This can lead to poor predictive performance when the link between moderator and effect size is continuous.

MetaCART, like any technique based on a single tree, inherits the aforementioned limitations. Moreover, the implications of metaCART's two-step approach, where cluster membership in the tree's terminal nodes is used as a categorical moderator in a subgroup meta-analysis, remain unexplored (as the authors rightly point out, Li et al., 2017). Several specific concerns regarding this approach are noteworthy: First, by treating cluster membership as an observed variable, metaCART ignores classification uncertainty: The possibility that a given study might be a close match for more than one of the tree's terminal nodes. The assumption that cluster membership is measured without error is untenable, because even slight changes in the unstable tree structure could lead studies to be assigned to a different terminal node (Hastie et al., 2009). Secondly, it is not clear that mixed-effects subgroup analysis is an appropriate model for the heterogeneity within and between groups resulting from a tree-based classification.

The mixed-effects model used by metaCART assumes heterogeneity to be normally distributed, and assumes the amount of heterogeneity within all subgroups to be the same (Viechtbauer & others, 2010). This is at odds with the fact that the non-parametric tree algorithm merely tries to maximize the differences between these clusters; it has no incentive to ensure that effect sizes within terminal nodes will be normally distributed, nor to ensure equality of variance (homoscedasticity) across terminal nodes. Thirdly, the mixed-effects subgroup analysis assumes a fixed effect across groups. In many cases, however, the random-effects model is more appropriate to model residual heterogeneity (Riley et al., 2011). Finally, using this subgroup analysis for statistical inference violates the logic of null-hypothesis significance testing. Given that the groups being compared are derived from a data-driven approach designed to maximize differences between them, the null-hypothesis that there are no differences between the groups is untenable. Conducting such a test is tautological at best (see: the null ritual, Gigerenzer, Krauss, & Vitouch, 2004).

Random Forests

Many of the limitations of single trees are overcome by the ensemble learning method "random forests" (Breiman, 2001). This aptly named technique grows many (often hundreds or thousands) of individual trees on bootstrapped samples of the initial dataset, which introduces variance between the trees. To increase this variance even further, at each split, only a small random selection of moderators is considered as splitting variables. The predictions of all these trees are then averaged. In doing so, random forests turn the instability of individual trees into an advantage: First of all, each variable has a chance to be included in at least some of the trees in the forest, and order effects are perturbed across trees. This allows for the detection of correlated predictors and smaller effects that would be washed out in a single tree. Secondly, because each tree captures some of the systematic patterns in the data, plus some of the idiosyncrasies present only in its bootstrap sample, overfitting tends to average out across trees. Thirdly, because each tree returns slightly different predictions, random forests capture continuous (curvi)linear relationships and complex interactions much

better than the piecewise linear predictions of individual trees can. Consequently, substantial gains in predictive accuracy can be achieved by using random forests rather than individual trees (Bühlmann & Yu, 2002). Finally, the predictive ability of each tree can be computed for the cases *not* part of that tree's bootstrap sample. This yields a measure of predictive accuracy known as *out-of-bag (OOB) error*, which closely approximates the cross-validation error, and thus provides a good estimate of the prediction accuracy in future samples (Boulesteix, Strobl, Augustin, & Daumer, 2008; Hastie et al., 2009). Consequently, the R^2_{oob} also approximates the cross-validated R^2_{cv}. In most random forests algorithms, weights are applied during bootstrap sampling, which means that cases with larger weights are selected with greater probability. Once studies are included in a bootstrap sample, however, each is weighted equally in the splitting process.

Interpreting random forests

Although random forests lack the intuitive appeal of single trees, they yield a useful metric for interpreting the overall importance of moderators, or "variable importance". The variable importance measure with the most desireable characteristics that is also widely implemented is *permutation importance* (Breiman, 2001, for an unbiased alternative see Altmann, Toloşi, Sander, & Lengauer, 2010). After growing the random forest, each variable is randomly scrambled, one at a time, to remove any association it may have had with the outcome. Permutation importance is then defined as the difference in OOB prediction error before- and after permutation. A negative permutation importance can be obtained when randomly permuting a variable leads to a spurious improvement in predictions. An advantage of permutation importance is that it takes into account the variable's contribution individually, as well as in interactions with other variables. Thus, if the permutation importance of a variable is high, but it does not appear to have a main effect on the outcome, this might indicate that the variable is relevant in interactions with other variables. The importance of all variables is typically plotted to compare their relative contributions, and can be used to identify and select important moderators. Secondly, random forests offer a way to

visualize the marginal effects of individual moderators on effect size, and to probe interactions. These so-called "partial dependence plots" display the predicted effect size at different levels of a specific moderator, whilst averaging over all other moderators. Because such plots can be easily visualized for one or two predictors, they lend themselves to exploring the functional form of the bivariate relationship between a moderator and effect size in a 2D plot, or to visualize the bivariate interaction between two moderators in a 3D plot, or heat map.

Developing MetaForest

The goal of the present study was to develop a robust, tree-based method to explore heterogeneity in meta-analysis. We aimed to overcome the limitations of single trees by using random forests. To this end, we developed MetaForest: A technique that applies random- or fixed-effects weights to random forests. We implemented MetaForest in R (Team, 2017), as a wrapper around `ranger` (Wright & Ziegler, 2015), which is a very fast implementation of the random forests algorithm with advanced features. The efficient implementation of ranger was instrumental in conducting the large-scale simulation studies presented. To optimize the speed of our simulation studies, we used a very simple implementation of MetaForest. A user-friendly version of MetaForest has been published in the R package `metaforest`, along with several functions to summarize and plot the results of analyses.

Estimating residual heterogeneity

To obtain random-effects weights, the residual variance, τ^2, must be known or estimated. In preliminary analyses, we considered three alternative approaches to this problem. The first was based on Li et al. (2017), who simply passed the "known" values of τ^2 used to generate the data to the metaCART function. A limitation of this approach is that it does not take into account simulation variance induced when random errors are drawn from a distribution with variance τ^2. Second, in their suggestions for future research, Li and colleagues suggested first growing an unweighted tree, and estimating τ^2 on the residuals of this analysis. Ostensibly, the goal of this approach is to remove the "reducible" heterogeneity from the data, leaving only the

"irreducible", residual heterogeneity. It is not clear, however, that an unweighted tree model can adequately account for the heterogeneity introduced by moderators. Finally, a third approach is to estimate τ^2 on the raw data. Because the raw data incorporates both residual heterogeneity, and heterogeneity introduced by the moderators, τ^2 will be overestimated if there are significant moderators. In the simulation studies presented below, we included all three of these estimates of τ^2. However, across both studies, the third estimate, based on the raw data, had the best performance. Because the differences between the three estimators were negligible, and because evaluating different estimators of τ^2 was not the focus of the present paper, we report only the results for this best-performing estimate. The full syntax for all analyses will be made available on the Open Science Framework (OSF). We used the method of moments (DerSimonian & Laird, 1986) for all estimates of residual heterogeneity, because it offers a substantial computational advantage over other, iterative, estimators.

Study 1

This first simulation constitutes a partial replication of the simulation Li, Dusseldorp, and Meulman (2017) conducted to determine which settings resulted in the best performance for metaCART. We included three versions of MetaForest: 1) random-effects weighted MetaForest, 2) fixed-effects weighted MetaForest, and 3) unweighted MetaForest. We included the version of metaCART recommended by the authors, which uses random-effects weights with a known value of τ^2, and controls overfitting by pruning the tree until cross-validation error is within the minimum cross-validation error plus $0.5SE$. We evaluated three performance criteria: 1) The algorithms' predictive performance, 2) their power, and 3) their ability to perform variable selection.

Performance criteria

Predictive performance. In order to examine the predictive performance of the different algorithms, we had to depart from the original study design. Li and colleagues simulated data using binary trees as the true data-generating models, and then defined performance as metaCART's ability to successfully retrieve these true

models. This metric does not generalize to evaluating the performance of other models, including random forests, nor to evaluating the performance of metaCART itself when the data-generating model is not a binary tree. Since meta-analysis is generally defined as a regression problem, we instead defined performance as prediction accuracy, operationalized as the cross-validated R^2_{cv} (Hastie et al., 2009). This is obtained by estimating a model on a "training" dataset, and then using a second "testing" dataset to calculate the fraction of variance explained by this model, relative to the variance around the mean of the training dataset, which is the best prediction for testing data in the absence of a model. Thus, for a testing dataset of n studies,

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}, \tag{5}$$

where $\hat{y}_i$ is the model prediction for study i, and $\bar{y}$ is the mean of the training dataset.

Power. Li and colleagues defined power as the algorithm's ability to correctly detect the presence of moderator effects. When an algorithm correctly detects moderator effects, its cross-validated R^2_{cv} is greater than 0. We thus defined power as the proportion of datasets in which an algorithm had an $R^2_{cv} > 0$.

Variable selection. Li and colleagues defined metaCART's ability to perform variable selection in terms of the algorithm's ability to exactly recover the underlying model. Again, this operationalization does not generalize to other algorithms which do not yield binary trees. Instead, we defined variable selection in terms of the algorithms' ability to assign positive variable importance values to relevant moderators. Both the single trees used by metaCART, and the random forests underlying MetaForest, yield variable importance measures, which capture the relative contribution of different moderators. We extracted these variable importance measures, and rescaled them to sum to 100 within each simulated dataset.

Design factors

Li and colleagues manipulated five design factors: the number of studies k (40, 80, and 120), the average within-study sample size $\bar{n}$ (40, 80, and 160), the number of moderators M (5, 10, and 20), the population effect size β (.5 and .8), and the residual

heterogeneity τ^2 (0, .025, and .05). The authors report omitting two conditions from the final paper, because metaCART displayed poor performance in a pilot simulation: A condition where the number of studies $k = 20$, and a condition where the population effect size $\beta = .2$. We chose to retain these conditions, to examine whether MetaForest might offer an advantage in dealing with small samples and effect sizes. Finally, data were simulated using a random-effects model (see Equation 2), based on five binary tree models $f(x)$: (a) a null-model without any moderators, i.e., $\mu_i = 0$; (b) a main effect of a single moderator, $\mu_i = \beta x_{1i}$; (c) a two-way interaction, $\mu_i = \beta x_{1i} x_{2i}$; (d) two two-way interactions, $\mu_i = \beta x_{1i} x_{2i} + \beta(1 - x_{1i}) x_{3i}$; and (e) a three-way interaction, $\mu_i = \beta x_{1i} x_{2i} x_{3i}$. In the original study, the null-model (a) was included to examine the metaCART's tendency to build a tree model when the data consisted only of noise. In the present study, such overfitting can be deduced from a negative $R^2_c v$. When no true model is present in the data, the best prediction an algorithm can make is the intercept of the training data. If an algorithm instead overfits the noise in the data, this will, on average, result in a negative R^2_{cv}.

Simulation

For every possible combination of the design factors and models, 100 meta-analytic datasets were simulated. Each dataset consisted of two subsamples: A training sample, with a number of studies equal to the design factor k, and a testing sample of 100 studies. For each study i, the values of M binary moderators were randomly drawn from a Bernoulli distribution with probability $p = .5$ (i.e., a series of coin tosses), resulting in the vector of moderators $\boldsymbol{x}_i$. The dependent variable y_i represents Hedges' g; an effect size commonly used in meta-analysis, which is an estimator of the standardized mean difference between a treatment and control group. For each study, we first sampled the true effect size θ_i from a normal distribution, with a mean computed by evaluating the model $f(x)$ for the value of β and the vector of predictors $\boldsymbol{x}_i$, and with variance τ^2 (the between-studies variance). Sampling error was introduced by varying the within-study sample size: Each study sample size n_i was drawn from a normal distribution with mean $\bar{n}$, and standard deviation $\bar{n}/3$

(Viechtbauer, 2007). The observed effect size y_i was then drawn from a non-central t-distribution, assuming an equal number of cases in the treatment and control group (Hedges, 1981; Li et al., 2017).

Hypotheses

Several hypotheses can be formulated about the influence of the design factors on the predictive performance of the algorithms. Broadly speaking, each design factor can be thought of as influencing either the "signal" in the data, or the "noise" that obscures this signal. Any design factor that increases the signal should have a positive effect, and any factor that increases the noise should have a negative effect on the variance explained by different algorithms. The effect size β, for example, largely determines the amount of signal in the data, and τ^2 determines the "noise" due to residual heterogeneity. The model determines the relative influence of the effect size β and τ^2, so these design factors are expected to interact with the model. The number of studies k constitutes the data available to the algorithms, from which to reconstruct the signal. The average $\bar{n}$ of these studies determines the quality of these data, as lower levels of $\bar{n}$ incur greater noise due to sampling error. Finally, the number of moderators M contributes indirectly to the noise in the data, because only a few of the moderators are correlated with the outcome, and the rest are irrelevant. Single trees are generally able to ignore uncorrelated predictors, and should thus incur little noise due to increasing M. Random forests, on the other hand, are forced to learn from uncorrelated moderators any time the random selection of variables available for a split does not include a moderator associated with the outcome (Hastie et al., 2009). Thus, MetaForest should be substantially more affected by uncorrelated moderators than metaCART.

We expected that MetaForest would demonstrate superior performance over metaCART, because of random forests' greater flexibility in learning from data. We further hypothesized that random-effects MetaForest would outperform fixed-effects and unweighted MetaForest, because the random-effects weights cause the algorithm to learn more from more informative studies.

Analyses

We examined the predictive performance of all algorithms, using ANOVAs to assess the influence of the design factors and their two-way interactions on both absolute and relative performance. Absolute performance was defined as the cross-validated R^2_{cv}. Relative performance was operationalized as the difference between the R^2_{cv} of random-effects MetaForest and that of the other algorithms, on the same dataset. We conducted separate ANOVAs for model (a) and the other models, because the predictors of overfitting should differ from the predictors of performance in the presence of a true effect. Due to the large size of the dataset (162000 observations), all effects were statistically significant, so we focus on effect size instead, in terms of the partial η^2. This measure reflects the proportion of the variance in the outcome explained by each independent variable, after all other effects have been partialed out. To examine the power of MetaForest versus metaCART, we examined the conditions under which the best-performing version of MetaForest, and metaCART, achieved a positive cross-validated R^2_{cv} in at least 80% of datasets.

Results

The results are visually presented in greater detail in Supplementary Figures S1-S5, which display the mean R^2_{cv} by algorithm, paneled by the design factors.

Overfitting the null model. For model (a), which did not include any predictors, the mean R^2_{cv} was negative for all versions of MetaForest; random-effects MetaForest $M = -0.07, SD = 0.07$, fixed-effects $M = -0.07, SD = 0.07$, unweighted $M = -0.08, SD = 0.08$. This indicates that all versions of MetaForest were susceptible to overfitting a null model. MetaCART on the other hand, which defaults to an intercept-only random-effects meta-analysis when no tree structure is found, was much less susceptible to overfiting ($M = 0.00, SD = 0.05$). According to ANOVA, two design factors played minor roles in predicting overfiting. The number of moderators M had an influence on all versions of MetaForest (η^2_ps = .06); overfitting decreased with an increasing number of moderators. In the absence of informative moderators, MetaForest overfits because it is forced to learn from uninformative moderators instead. This

finding suggests that, when there are more uninformative moderators to choose from at each split, this effect is diluted. Secondly, the number of studies k had an effect on overfitting for all algorithms (η_p^2s ranged from .01 – .02), and overfitting decreased as k increased.

Absolute performance. For the remaining models, the mean R_{cv}^2 was highest for random-effects MetaForest ($M = 0.18, SD = 0.21$) and fixed-effects MetaForest ($M = 0.18, SD = 0.21$), followed by unweighted MetaForest ($M = 0.17, SD = 0.22$), and finally metaCART ($M = 0.16, SD = 0.24$). The partial η^2s of the effects of design factors on predictive performance are displayed in Table 1. Because many factors had small, non-zero effects for some of the algorithms, we limit our discussion here to effects with a partial $\eta^2 \geq .10$ for at least one of the algorithms.

Across algorithms, we found large effects of the effect size β, and the interaction between β and the model (see Figure 1). These effects were strongest for the different versions of MetaForest, and smaller for metaCART. Predictably, the positive effect of β on predictive performance was lower for models in which β was multiplied with more moderators. The residual variance τ^2 showed a main effect, but did not interact with the model. This main effect indicated that increasing τ^2 diminished predictive performance of MetaForest, and to a lesser extent, metaCART (Figure 6).

All algorithms showed substantial effects of the number of studies k. Moreover, the interaction between k and β played a substantial role for metaCART, and a small role for MetaForest (Figure 3). Specifically, adding additional studies had a stronger effect on predictive performance when the effect size was greater. For metaCART only, there was also a substantial interaction effect between k and the model (Figure 4), which revealed that, as model complexity increases, metaCART needs a larger sample size before it starts explaining any variance. The different versions of MetaForest did not show this effect. There was also a substantial main effect of the average within-study sample size: Greater $\bar{n}$ substantially improved predictive performance for MetaForest, and to a lesser extent, for metaCART (Figure 5). This reflects the positive influence of the information added by having large sample studies, which yield more

precise observed effect sizes. Finally, the number of moderators M and its interaction with β played a substantial role for MetaForest, but not for metaCART. As can be seen in Figure 2, for MetaForest, the positive effect of β on predictive performance was diminished by increasing the number of moderators. This reflects the fact that MetaForest is forced to learn from irrelevant predictors, if these outnumber true moderators.

Relative performance. On average, the mean performance difference between random-effects MetaForest and metaCART was $M = 0.02, SD = 0.16$. The differences with fixed-effects and unweighted MetaForest were $M = 0.001, SD = 0.018$ and $M = 0.01, SD = 0.03$, respectively. These results indicate that random-effects MetaForest has a small advantage over all other algorithms. None of the design factors exerted notable influence on the performance difference between random-effects MetaForest, and fixed-effects or unweighted MetaForest. These absolute performance differences between the different versions of MetaForest thus appear to be relatively constant with respect to the conditions of this simulation. Because these differences were small and constant, we will focus only on the comparison between random-effects MetaForest and metaCART going forward.

Several design factors exerted substantial influence on the difference between random-effects MetaForest and metaCART. Table 1 shows the partial η^2s of these effects. Most relevant are two interaction effects: The interaction between k and the effect size β, and the interaction between the number of studies k and the model. Inspection of the interaction between k and β suggests that the effect of the number of studies on the performance difference is greater at higher effect sizes (see Figure 3). Specifically, MetaForest performs relatively well, even when sample size is as low as $k = 20$, when the effect size is at least $\beta = 0.5$, but at larger sample sizes of $k \geq 80$, metaCART outperforms MetaForest. Inspection of the second interaction, between k and the model (Figure 4), suggests that MetaForest already performs well at low levels of k, and that metaCART needs k to surpass a certain threshold before it starts performing well. This small-sample advantage of MetaForest increases with the

complexity of the model.

Power. To determine the statistical power of random-effects MetaForest and metaCART, we computed the proportion of cases for each cell of the design in which each algorithm achieved a positive cross-validated R^2_{cv}, that is, produced superior predictions as compared to the mean of the training dataset. These results are reported in full in supplementary Table S1. To determine practical guidelines for the power of these algorithms, we focus here on the conditions under which the two algorithms achieved a positive cross-validated R^2_{cv} in at least 80% of datasets (see Figure 7). The results reflect the influence of the design factors as discussed in the preceding paragraphs, and demonstrate that random-effects MetaForest reaches sufficient power at lower levels of k and $\bar{n}$ than metaCART. When the effect size is small ($\beta = .2$), both algorithms have insufficient power across most conditions. Only for relatively simple models (b, c, and d), and with relatively high levels of $\bar{n}$ and k, do the algorithms reach acceptable power. For medium and large effect sizes ($\beta \in \{.5, .8\}$), MetaForest has sufficient power even at the lowest levels of k (20) and $\bar{n}$ (40) for models b, c, and d, and metaCART has sufficient power at higher levels of k (Li and colleagues recommend at least $k = 80$). For the more complex model (e), both algorithms require higher levels of k and $\bar{n}$: Approximately $k = 80$ is required for MetaForest to reach sufficient power, and $k = 120$ for metaCART.

Variable selection. We conducted ANOVAs to examine which design factors predicted the standardized variable importance of a relevant moderator (we examined x_1). The most important design factors were effect size β (partial η^2 .22 and .18 for random-effects MetaForest and metaCART), the number of studies k (partial η^2 .03 and .18 for random-effects MetaForest and metaCART), and the interaction between these design factors (partial η^2 .02 and .13 for random-effects MetaForest and metaCART). Figures 8 and 9 represent the distribution of the standardized variable importance for the first 5 variables in the model (this includes all relevant moderators, and at least one irrelevant moderator for comparison), paneled by β, k, and the model. Each segment of a boxplot (whiskers and both halves of the box) represents 25% of scores. Thus, when a

boxplot is raised above the x-axis, an algorithm identifies a specific variable as having positive importance at least 75% of the time. Using this cutoff, it can be seen that metaCART has adequate power for detecting relevant moderators only when the effect sizes is at least $\beta \geq .5$, and the number of studies is at least $k \geq 40$ in model (b), $k \geq 80$ in models c and d, and $\beta = .8$ and $k = 120$ for model (e). MetaForest, on the other hand, had adequate power in all conditions for model (b). For models c and d, MetaForest had sufficient power in all conditions if $\beta > .2$. When $\beta = .2$, MetaForest required at least $k \geq 80$ studies. For model (e), MetaForest had sufficient power if $\beta \geq .5$ and $k \geq 40$. Thus, MetaForest displayed a clear advantage in detecting relevant moderators in most conditions.

Discussion

Overall, random-effects MetaForest outperformed metaCART in terms of predictive performance, power, and variable selection. This study yielded three important insights regarding MetaForest: First, it revealed that the performance advantage of random-effects MetaForest over the other versions of MetaForest was negligible, reminiscent of the minimal difference between unweighted and random-effects metaCART reported by Li, Dusseldorp, and Meulman (2017). Secondly, it illustrated that MetaForest inherits the tendency of random forests to overfit in the absence of a true model, and in the presence of many irrelevant moderators. This problem is mitigated by the fact that MetaForest provides an estimate of the R^2_{oob}, which indicates when the model is overfitting by decreasing or becoming negative with the inclusion of noise predictors. MetaCART did not suffer from the same sensitivity to irrelevant moderators. Thirdly, it demonstrated MetaForest's superior power, when compared with metaCART, to detect small effects and complex models at smaller sample sizes.

Study 2

Li, Dusseldorp, and Meulman (2017) point out that future research is required to examine metaCART's performance in cases where the data-generating mechanism is not a binary tree. This is a relevant question, because — except for the obvious example of biological sex — true binary data-generating processes are exceedingly rare in the social

sciences. Even when moderators are *measured* as binary variables, such as the presence or absence of behavior change techniques that Li and colleagues examined, the underlying distribution is often continuous: For example, interventions differ not merely in terms of whether or not they included a specific behavior change technique, but also in terms of the quality and quantity of implementation. In this second simulation study, we thus made the more common assumption that moderators are continuous and normally distributed. Normal distributions are ubiquitous in nature, and in the social sciences, because they are the maximum entropy distribution out of all potential distributions with mean μ and variance σ^2, meaning that they require the least assumptions about the underlying data-generating mechanism (Lyon, 2013). Secondly, in most conditions of Study 1, the number of uncorrelated moderators outnumbered the number of true moderators substantially. It is debatable whether it is realistic to assume that researchers examining heterogeneity might include 20 irrelevant study characteristics for every relevant one. At the very least, this problem can be avoided by screening potential moderators before including them in the analysis. Thus, following Higgins and Thompson (2004), we instead used a range of 1-5 uncorrelated moderators. We kept the number of uncorrelated moderators constant across datasets by generating the number of moderators required for the true model, plus the specified number of uncorrelated moderators. Finally, the range of residual heterogeneity examined by Li and colleagues ($\tau^2 \in [0, .05]$) may have been slightly conservative. Other simulation studies have used larger ranges, such as $\tau^2 \in [0, 1]$ (Viechtbauer, 2007), or even $\tau^2 \in [0, 5]$ (Higgins & Thompson, 2004). A recent open-data study of 705 meta-analyses published in *Psychological Bulletin* from 1990-2013 (Van Erp, Verhagen, Grasman, & Wagenmakers, 2017) found that, for measures of mean differences, the reported heterogeneity ranged from $[0, 1]$. However, as these data were extremely positively skewed, we used the minimum, median, and 90th percentile of this distribution, corresponding to the values 0, .04, and .28, to obtain a range of τ^2 as encountered in practice.

The design factors of this second study were the number of studies k (40, 80, and

120), the average within-study sample size $\bar{n}$ (40, 80, and 160), the number of uncorrelated moderators M (1, 2, and 5), the population effect size β (.2, .5, and .8), and the residual heterogeneity τ^2 (0, .5, and 1). Data were simulated using the random-effects model (see Equation 2), based on five linear models $f(x)$:

(a) main effect of one moderator, $\mu_i = \beta x_{1i}$

(b) two-way interaction, $\mu_i = \beta x_{1i} + \beta x_{2i} + \beta x_{1i} x_{2i}$

(c) three-way interaction,
$\mu_i = \beta x_{1i} + \beta x_{2i} + \beta x_{3i} + \beta x_{1i} x_{2i} + \beta x_{1i} x_{3i} + \beta x_{2i} x_{3i} + \beta x_{1i} x_{2i} x_{3i}$

(d) two two-way interactions, $\mu_i = \beta x_{1i} + \beta x_{2i} + \beta x_{3i} + \beta x_{4i} + \beta x_{1i} x_{2i} + \beta x_{3i} x_{4i}$

(e) non-linear, cubic relationship, $\mu_i = \beta x_{1i} + \beta x_{1i}^2 + \beta x_{1i}^3$.

Simulation and analyses

The simulation was identical to Study 1, with the exception that the moderators for each simulated dataset (equal to M plus the number of moderators in the model) were drawn from a standard normal distribution. The hypotheses and analyses were the same as in Study 1.

Results

The results are visually presented in greater detail in Supplementary Figures S6-S10, which display the mean R^2_{cv} by algorithm, paneled by the design factors.

Absolute performance. The mean R^2_{cv} was highest for random-effects MetaForest ($M = 0.38, SD = 0.22$), followed by Uniform MetaForest ($M = 0.38, SD = 0.22$), and fixed-effects MetaForest ($M = 0.35, SD = 0.20$), and finally metaCART ($M = 0.20, SD = 0.27$). The partial η^2s of the effects of design factors on predictive performance are displayed in Table 2.

For all algorithms, we found large effects of β and τ^2. The interactions between β and the model were much smaller than in Study 1 (see Figure 11). The interactions between τ^2 and the model, however, were larger (see Figure 11). Specifically, predictive performance increased with increasing effect size, and decreased with increasing τ^2, and

the steepness of this change differed between models. All of these effects were strongest for random-effects MetaForest, followed by unweighted MetaForest, and were much smaller for metaCART.

The number of studies k also had a substantial effect on the performance of all three algorithms. Moreover, for metaCART, and to a much smaller extent, for MetaForest, the number of studies k interacted with the residual variance τ^2, and with the model. The interaction between k and the residual variance (Figure 12) suggests that metaCART, in particular, benefits less from additional studies k when the residual heterogeneity is higher. The interaction between k and the model (Figure 13) revealed that, for metaCART, a minimum number of studies was required before the model started explaining variance, and this threshold differed between the different models.

The average within-study sample size $\bar{n}$ had a much smaller effects in this study, compared to Study 1, all $\eta_p^2 \leq .06$. Exploratory inspection of the marginal effect of $\bar{n}$ revealed that, especially for random-effect and unweighted MetaForest, greater average within-study sample sizes were associated with better predictive performance (see Figure 14). The effect of the number of uncorrelated moderators M was smaller as well, and revealed that a larger number of uncorrelated moderators still led to worse performance (see Figure 15) for all versions of MetaForest, but not for metaCART. This illustrates that MetaForest learns more from additional information than metaCART, but remains vulnerable to the presence of irrelevant moderators.

Relative performance. On average, the mean performance difference between random-effects MetaForest and metaCART was $M = 0.18, SD = 0.19$. The differences with fixed-effects and unweighted MetaForest were $M = 0.03, SD = 0.07$ and $M = 0.001, SD = 0.034$, respectively. These numbers suggest a substantial advantage of random-effects MetaForest over metaCART, and a very small advantage of random-effects MetaForest over fixed-effects and unweighted MetaForest. None of the design factors predicted the performance difference between random-effects and fixed-effects MetaForest, but several predicted the difference between random-effects MetaForest, and unweighted MetaForest and metaCART. Table 2 shows the partial η^2s

of the effects of design factors on relative performance.

For the difference between random-effects and unweighted MetaForest, the most important predictors were the model, effect size β, and the interaction between these factors (see Figure 11). Inspection of this interaction revealed that random-effects MetaForest slightly outperformed unweighted MetaForest, except when effect size was large, in models with a single relevant moderator (models a and e). Smaller effects for the number of studies k, and the interaction between k and the model, also suggested that random-effects MetaForest slightly outperformed unweighted MetaForest, except when the number of studies was large, in models with a single relevant moderator (see Figure 13). Together, these findings suggests that, when the "signal" in the data is strong (i.e., large effect size and/or number of studies), and the model is simple, random-effects weighting stops being beneficial, and instead introduces unnecessary noise.

Regarding the difference between random-effects MetaForest and metaCART, the number of studies k, the model, and the interaction between these factors had substantial effects (Figure 13). This interaction revealed that MetaForest already performs well at low levels of k, whereas metaCART needs a minimum number of studies before it starts performing well. This small-sample advantage of MetaForest was especially notable for more complex models. Another relevant factor was the effect size β (for example, see Figure 11): Compared to metaCART, the performance of MetaForest improved more strongly with increasing β.

Power. The power of random-effects MetaForest and metaCART was determined by computing the proportion of cases in which each algorithm achieved a positive cross-validated R^2_{cv} in each cell of the simulation design. These results are reported in full in supplementary Table S2. To determine practical guidelines for the statistical power of these algorithms, we examined the conditions under which random-effects MetaForest and metaCART, achieved a positive cross-validated R^2_{cv} in at least 80% of datasets (see Figure 16). The results reflect the influence of the design factors as discussed in the preceding paragraphs, and show that MetaForest had

sufficient power in most conditions, even for as little as $k = 20$ studies, except when the effect size was small ($\beta = 0.2$), and residual heterogeneity was high ($\tau^2 = 0.28$). MetaCART, by contrast, required upward of $k = 40$ studies for the univariate models a and e, and upward of $k = 80$ for models with bivariate interactions, b and d.

The pattern of results was more nuanced when the effect size was small and residual heterogeneity was high. In this case, neither algorithm reached sufficient power for model (a). For models b and c, MetaForest required about $k = 80$ studies, and metaCART did not reach sufficient power. For model (d), MetaForest required only $k = 20$ studies in most cases, and metaCART did not reach sufficient power. For model (e), MetaForest reached sufficient power at $k = 20$, and metaCART at $k = 40$.

Variable selection. We conducted ANOVAs to examine which design factors predicted the standardized variable importance of a relevant moderator (we examined x_1). The most important design factors for MetaForest and metaCART were effect size β (partial η^2s .12 and .11, respectively), the model (partial η^2s .19 and .17, respectively), and the interaction between these factors (partial η^2s .13 and .11, respectively). The number of studies k was also relevant, particularly for metaCART (partial η^2s .02 and .13 for MetaForest and metaCART), and the interaction between k and the model (partial η^2s .02 and .12 for MetaForest and metaCART).

Figures 17 and 18 illustrate the distribution of the standardized variable importance for the first 5 variables in the model (this includes all relevant moderators, and at least one irrelevant moderator for comparison), paneled by β, k, and the model. When a boxplot is raised above the x-axis, an algorithm identifies a specific variable as having positive importance at least 75% of the time. Using this cutoff, it can be seen that MetaForest has adequate power for detecting relevant moderators in all conditions, except when $k = 20$. At the lowest level of k, MetaForest had adequate power for univariate models a and e, low power when the effect size was small ($\beta = 0.2$) for models b and d, with bivariate interactions, and low power across all effect sizes for model (c), with a three-way interaction. MetaCART, on the other hand, never reached adequate power to detect moderators at $k = 20$. For model (a), it reached adequate power if

$k \geq 40$ and $\beta \geq .5$; for model (b), it reached sufficient power if $k \geq 80$ and $\beta \geq .5$, and for all effect sizes if $k = 120$. For three-way interaction model (c), metaCART never reached adequate power. For model (d), metaCART's power was satisfactory if $k \geq 120$ and $\beta \geq .5$, or if $k \geq 80$ and $\beta = .8$. Finally, for model (e), metaCART reached adequate power in all conditions, if $k \geq 40$. These results illustrate that, across the board, MetaForest has greater power to detect moderators, and that the performance of metaCART is more strongly affected by the different design factors.

Discussion

Overall, random-effects MetaForest demonstrated the best predictive performance, closely followed by unweighted and fixed-effects MetaForest, and finally metaCART. The performance difference between random-effects and unweighted MetaForest was largely explained by the effect size and the model. Specifically, unweighted MetaForest outperformed random-effects weighted MetaForest when the effect size was large, and when the model included a single moderator. This lapse in performance is likely caused by an overestimation of the residual variance. Recall that there are two sources of between-studies heterogeneity in the data: On the one hand, heterogeneity is caused by the true data-generating model, composed of the effect size and the model. On the other hand, random, residual heterogeneity is added (τ^2). Because MetaForest estimates residual heterogeneity from the raw data, without accounting for the influence of the moderators, it will be overestimated. The more heterogeneity is introduced by the true data-generating model, the larger this over-estimation.

Nevertheless, random-effects MetaForest demonstrated consistently high predictive performance, and responded most flexibly to changes in the design factors. It also substantially outperformed metaCART, especially when the number of studies was low. Just as in Study 1, the performance of MetaForest still showed a slight decrease when the number of irrelevant moderators increased, but this effect was small. Moreover, random-effects MetaForest demonstrated superior power over metaCART in terms of detecting reliable patterns in the data, and variable selection.

General discussion

The present paper introduced MetaForest, a method for exploring heterogeneity in meta-analysis using random forests. In two simulation studies, we examined the predictive performance of random-effects, fixed-effects, and unweighted MetaForest, as well as the single tree-based technique metaCART (Li et al., 2017). Across both studies, random-effects MetaForest demonstrated superior predictive performance to metaCART. In most conditions, random-effects MetaForest slightly outperformed fixed-effects and unweighted MetaForest. These findings indicate that random-effects MetaForest is a promising technique for exploring heterogeneity. Classic approaches to meta-analysis are only suitable when all studies are similar, or when a limited number of known moderators explains between-studies heterogeneity, and the ratio of studies to moderators is high. MetaForest provides a solution when the number of potential moderators to be considered is large relative to the number of studies, or when no clear theory is available to limit the number of potential moderators. In such cases, MetaForest has greater statistical power than the classic approaches, and can be used to identify the most important moderators from a large number of candidates.

MetaForest demonstrated a performance advantage over metaCART in Study 1, and more so in Study 2. The greater performance advantage in Study 2 appears to be a consequence of two design choices: First, in Study 1, the data were generated from binary tree models, which can be exactly reproduced by metaCART, but not by MetaForest. Study 2 instead used continuous moderators. It is open to debate which data-generating model more accurately represent realistic research scenarios, although we have argued that normal distributions are likely a better approximation. Secondly, Study 1 included large numbers of irrelevant moderators, which are known to cause overfitting in random forests (Hastie et al., 2009). This illustrates that researchers should be mindful not to include "everything but the kitchen sink" as a moderator in MetaForest analyses. Even data-driven techniques require making informed decisions. The danger of overfitting is mitigated, however, by the fact that MetaForest's R^2_{oob} approximates the R^2_{cv} (if the data are independent). When the addition of moderators

leads to a decreasing or negative R^2_{oob}, this suggests that some of the moderators contribute nothing but noise, which the algorithm is overfitting. In this case, a researcher might want to perform some type of preliminary feature selection (e.g., Svetnik, Liaw, Tong, & Wang, 2004).

The present study suggested that MetaForest has substantial power for identifying relevant moderators. Assuming continuously distributed moderators, a mere 20 studies sufficed for MetaForest to attain acceptable power under most conditions, except when effect size was small, and residual heterogeneity was high. Researchers should be wary, however, of including large numbers of potentially irrelevant moderators, because these can undermine the algorithm's performance. MetaForest also has several advantages over single trees: It has greater power, is able to make smoother predictions, gives an estimate of the cross-validation error through its OOB error, and yields useful measures of variable importance and partial prediction plots. Although we used Hedges' g as the simulated effect size, the technique should generalize to other measures of effect size, just like classic meta-analysis (Viechtbauer & others, 2010).

Strengths and future directions

The present paper has several strengths. First, we were able to demonstrate that MetaForest had superior predictive performance and power over metaCART, even in a close replication of the simulation study used to validate metaCART, whose design factors placed MetaForest at a disadvantage. Secondly, we addressed one of the directions for future research suggested by Li, Dusseldorp, and Meulman (2017) by examining the performance of both algorithms in the presence of normally distributed moderators. Moreover, this second simulation used realistic estimates of τ^2, based on data from 705 published psychological meta-analyses (Van Erp et al., 2017).

Several limitations remain to be addressed in future research, however. The most important issue is how to handle dependent data in MetaForest. Oftentimes, studies report several effect sizes; for example, if a study examines the same research question using two different outcome variables, this study offers two effect sizes for meta-analysis. When such dependent effect sizes are more similar than effect sizes from independent

studies, they can introduce bias in meta-analyses. In classic meta-analysis, this problem arises because studies with several effect sizes incur more weight in the calculation of the summary effect (Borenstein et al., 2009). In random forests, predictions for new data are relatively robust against dependent data (Karpievitch, Hill, Leclerc, Dabney, & Almeida, 2009), but bias is introduced when computing the OOB (out-of-bag) prediction error. This bias extends to measures derived from the OOB error, such as the model's R^2_{OOB}, and the permutation importance. Recall that trees are grown on bootstrapped samples, and for each tree, the OOB error is calculated across cases not included in this bootstrapped sample. If the OOB sample contains effect sizes from the same studies included in the bootstrapped sample, the model will be able to predict these dependent effect sizes better than independent effect sizes from new studies, causing it to underestimate the OOB error. This problem becomes exacerbated if the similarity of effect sizes within studies increases (Karpievitch et al., 2009).

There are some solutions to the problem of dependent data. The simplest solution is to use only one effect size per study, either selected randomly, based on quality criteria, or fit to the research question. The obvious downside is that this does not make use of all available information. A second potential solution is to split the dataset into two cross-validation samples *by study*, thus including all dependent effect sizes from each study in the same cross-validation sample. Then, two random forests can be grown on these cross-validation samples, and for each random forest, the other sample can be used to calculate prediction error and variable importance (see Janitza, Celik, & Boulesteix, 2016). We have implemented this functionality in MetaForest, and plan to conduct a simulation study to evaluate this approach as a solution to the problem of dependent data. Finally, a method has been develped to conduct "multilevel" random forests (Hajjem, Bellavance, & Larocque, 2014). This might prove to be a very elegant solution to the problem of dependent data, but the approach is still very new, and additional research on this issue is required.

A second issue that might benefit from further exploration is the estimation of residual heterogeneity. In preliminary analyses, we included three different estimates of

residual heterogeneity: 1) The "known" value of τ^2 used for the simulation, which is subject to simulation variance; 2) an estimate of τ^2, based on the residuals of an unweighted MetaForest model; and 3) an estimate of τ^2 based on the raw data. We found an overall advantage for this latter estimate, and used it for the analyses reported in this paper. Although this was not a focus of the paper, we observed that the "known" value of τ^2 performed the worst, possibly because this estimate does not take simulation variance into account. The values estimates of τ^2 based on the residuals of an unweighted MetaForest analysis substantially underestimated true residual heterogeneity, which might suggest that this estimate is negatively biased. Finally, the estimate based on the raw data displayed the best performance on average. However, when the effect size was large, and the model involved a single predictor, unweighted MetaForest demonstrated superior performance over random-effects MetaForest. As mentioned before, this is likely due to an over-estimation of τ^2, which, when computed on the raw data, is inflated by moderator-induced heterogeneity. Future research should address the question of estimating τ^2 in tree-based meta-analysis more thoroughly.

It should also be noted that the performance differences between MetaForest's weighting schemes were minimal. One potential explanation for this might be that MetaForest is already so efficient at identifying reliable patterns in the data, that the additional advantages conferred by the different weighting schemes are marginal. Intriguingly, Li and colleagues similarly reported that the performance difference between the different weighting schemes was negligible for metaCART; evident only at the third decimal place. It is unclear whether these findings are comparable to ours, however, because Li and colleagues used a different performance metric than we did, and because weights are implemented differently in metaCART and MetaForest: In metaCART, weights are implemented in the impurity function used to evaluate the homogeneity of the groups resulting from a potential split. The influence of weights on model building is thus direct: Studies with smaller weights exert less influence in determining what splits are made. In MetaForest, by contrast, weights are implemented during bootstrap sampling. Studies with larger weights have a greater chance of being

included in each bootstrap sample. Once these studies are included, however, their influence on the splitting process is the same as that of any other study. Consequently, in bootstrap weighting, the weights exert only indirect influence on the model building process. Future research might implement both weighting approaches in MetaForest, and examine whether impurity weights are more advantageous than weighted bootstrapping.

There are also several minor limitations. For example, future studies might develop the design of the simulation study further. Although we addressed an important limitation of the prior literature by including continuous, rather than binary, moderators, these moderators were uncorrelated. Future research should examine the performance of tree-based approaches to meta-analysis in the presence of varying levels of multicollinearity amongst moderators. Furthermore, the performance of random forests is known to be responsive to two tuning parameters: The number of candidate variables considered for each split, and the number of trees in the forest (Hastie et al., 2009). Applied researchers are advised to evaluate the effects of varying these parameters in their analyses (e.g., see Strobl et al., 2009).

Recommendations for applied research

As we mentioned in the introduction, MetaForest aims to address the issue that arises when between-studies heterogeneity is suspected, but the specific moderators responsible for heterogeneity are unknown. To help applied researchers get started with this new technique, we have published the R package `metaforest`, and offer several recommendations for its use. The first recommendation precedes analysis, and relates to the design of the meta-analysis. When the search for moderators is exploratory, researchers ought to be inclusive, but focus on moderators that are expected to be relevant, including theoretically relevant moderators, as well as moderators pertaining to the sample, methods, instruments, study quality, and publication type. In our experience, many applied researchers code such study characteristics anyway, but omit them from their analyses for lack of statistical power. Each moderator should be coded either continuously or categorically. Secondly, missing data must be accounted for.

Ideally, missing data should be avoided; for example, by contacting authors, or finding the relevant information in other publications on the same dataset. Alternatively, single imputation must be used, because no method currently exists for aggregating the results of MetaForest models conducted on multiply imputed datasets. Finally, when all data are collected, the effect sizes must be computed, and their variance estimated. The R package `metafor` is well suited for this task (Viechtbauer & others, 2010). The resulting effect sizes and their variances, as well as the moderators that are to be included in the model, can then be passed to the MetaForest function.

With regard to data analysis, we recommend the use of random-effects weighted MetaForest by default, because it demonstrated the best performance in most conditions of our simulations. When reporting their results, researchers should substantiate their decision to explore heterogeneity on both subjective and objective grounds. The former can be achieved by simply ascertaining that the body of literature to be meta-analyzed appears to be heterogeneous; the same rationale commonly used to support the use of random-effects meta-analysis (Higgins et al., 2009). The latter can be accomplished by conducting a random-effects meta-analysis without any moderators, and reporting the estimated τ^2, or another estimate of heterogeneity. Cochran's Q-test for the significance of this heterogeneity can also be reported. It does not constitute sufficient grounds, however, for deciding whether to explore or ignore heterogeneity, because it is often underpowered when the number of studies is low, and overpowered when it is high (see Higgins & Thompson, 2002). Researchers should report the R^2_{oob}, to determine whether the MetaForest model explains significant variance in the effect size. To determine which moderators most strongly explain differences in effect size, variable permutation importance can be plotted and reported. The marginal effects of these moderators, and bivariate interactions between them, can be explored using partial prediction plots. The R package `metaforest` provides convenient functions to generate these statistics and plots.

Although MetaForest provides a wealth of fine-grained information about the relationships between moderators and effect size, we understand that applied

researchers are typically more comfortable with the regression-based paradigm of classic meta-analysis. As we pointed out in the introduction, there are potential pitfalls to consider when combining these two approaches. Nevertheless, we believe that a compromise is possible. Researchers might wish to follow a MetaForest analysis with a classic meta-regression analysis, to examine the effects of the most important predictors. Meta-regression provides a much cruder model than MetaForest, because it can only account for linear effects, whereas MetaForest can accommodate non-linear effects and complex interactions. However, it is an empirical question how much information is lost through this simplification. Researchers can do two things to determine whether this simplification is acceptable: First, researchers should examine partial prediction plots to determine whether the effects appear to be linear, or if predictors should be transformed to create non-linear- and interaction terms, before conducting meta-regression. Secondly, the cross-validated R^2_{cv} of the meta-regression model should be compared to the R^2_{oob} of the MetaForest model. If the difference between the two is not too large, the meta-regression model might serve as a viable and readily interpretable approximation of the effects of the most important moderators on effect size. As we pointed out in the introduction, one should be skeptical of statistical inferences drawn from a meta-regression which was preceded by variable selection, and focus on interpreting the direction and effect size of regression slopes, rather than null-hypothesis significance testing.

We expect that researchers might also be interested in comparing the performance of MetaForest and metaCART on their own data. To do so, we recommend using cross-validation to compare the R^2_{cv} for both algorithms. To evaluate the performance of both algorithms on large datasets, researchers can simply set aside a random subsample to compute the R^2_{cv}, as we did in the present study. For smaller datasets, researchers might instead prefer to use k-fold or leave-one-out cross-validation (Hastie et al., 2009). The same logic we proposed for comparing MetaForest models to meta-regression models applies here: As MetaForest is more flexible, its R^2_{cv} should be higher, and can serve as a benchmark for the performance that can be attained using metaCART. If the

R^2_{cv} of metaCART is only slightly lower, a single tree may be already be a relatively good — albeit simplified — approximation of the underlying model. However, it is important to keep in mind that the apparent interpretability of the structure of single trees can be misleading with regard to variable importance, and the true model. Variable importance measures should therefore be obtained using MetaForest, and MetaForest's partial prediction plots can be used to explore apparent interactions gleaned from metaCART's tree structure. Finally, it is important to keep in mind that drawing inferences from a null-hypothesis significance test of the difference between clusters derived from a data-driven procedure is problematic. When interpreting the subgroup analysis, we therefore recommend focusing on the *effect size* of the difference between clusters, rather than their significance.

Finally, with regards to publication, we highly recommend sharing the data and syntax for the meta-analysis publicly, either as supplementary materials, or on OSF, or by publishing the data (e.g., in the Journal of Open Psychology Data). This serves dual purposes: First of all, transparancy is likely to inspire confidence in the use of new techniques, which can be checked and replicated by readers and reviewers alike. Secondly, MetaForest can be used by future researchers to obtain fine-grained predictions of the expected effect size for a new study on the same topic, for example, in order to conduct power analysis. To this end, researchers can simply enter their planned design (or several alternative designs) as new lines of data, using the codebook of the original meta-analysis, and use the published MetaForest model to calculate the predicted effect size for a study with these specifications.

Conclusion

The present research has demonstrated that MetaForest is a powerful tool for exploring heterogeneity in meta-analysis, with a number of advantages over alternative tools. It can identify important moderators from a larger set of potential candidates, even when the number of studies is low. If moderators are continuously distributed, MetaForest often has sufficient power with as little as 20 studies. This is an appealing quality, because many meta-analyses have small sample sizes. Moreover, MetaForest

yields a measure of variable importance which can be used to identify important moderators, and offers partial prediction plots to explore the shape of the marginal relationship between moderators and effect size. Although tree-based approaches such as MetaForest constitute a fully fledged, comprehensive paradigm for data analysis, they can also be readily integrated in classical meta-analytic approaches: If MetaForest is conducted as a primary analysis, classic meta-analysis can be used to quantify heterogeneity, and to provide a simplified representation of the linear effects of important predictors. Conversely, a theory-driven classical meta-analysis could be complemented by MetaForest, as a final check to ensure that important moderators have not been overlooked. We hope that this approach will be of use to applied researchers, and that the availability of user-friendly R functions will facilitate its adoption.

References

Altmann, A., Toloşi, L., Sander, O., & Lengauer, T. (2010, May). Permutation importance: A corrected feature importance measure. *Bioinformatics*, *26*(10), 1340–1347. doi: 10.1093/bioinformatics/btq134

Borenstein, M., Hedges, L. V., Higgins, J. P. T., & Rothstein, H. R. (2009). *Introduction to Meta-Analysis*. John Wiley & Sons, Ltd. doi: 10.1002/9780470743386

Boulesteix, A.-L., Strobl, C., Augustin, T., & Daumer, M. (2008, February). Evaluating Microarray-based Classifiers: An Overview. *Cancer Informatics*, *6*, 77–97.

Braver, S. L., Thoemmes, F. J., & Rosenthal, R. (2014, May). Continuously Cumulating Meta-Analysis and Replicability. *Perspectives on Psychological Science*, *9*(3), 333–342. doi: 10.1177/1745691614529796

Breiman, L. (2001, October). Random Forests. *Machine Learning*, *45*(1), 5–32. doi: 10.1023/A:1010933404324

Bühlmann, P., & Yu, B. (2002). Analyzing Bagging. *The Annals of Statistics*, *30*(4), 927–961. doi: 10.2307/1558692

Cesario, J. (2014, January). Priming, Replication, and the Hardest Science. *Perspectives on Psychological Science*, *9*(1), 40–48. doi: 10.1177/1745691613513470

DerSimonian, R., & Laird, N. (1986, September). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188. doi: 10.1016/0197-2456(86)90046-2

Dusseldorp, E., van Genugten, L., van Buuren, S., Verheijden, M. W., & van Empelen, P. (2014). Combinations of techniques that effectively change health behavior: Evidence from Meta-CART analysis. *Health Psychology*, *33*(12), 1530–1540. doi: 10.1037/hea0000018

Earp, B. D., & Trafimow, D. (2015, May). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in Psychology*, *6*. doi: 10.3389/fpsyg.2015.00621

Fabrigar, L. R., & Wegener, D. T. (2016, September). Conceptualizing and evaluating the replication of research results. *Journal of Experimental Social Psychology*, *66*,

68–80. doi: 10.1016/j.jesp.2015.07.009

Gigerenzer, G., Krauss, S., & Vitouch, O. (2004). The null ritual : What you always wanted to know about significance testing but were afraid to ask. In D. Kaplan (Ed.), *The Sage handbook of quantitative methodology for the social sciences* (pp. 391–408). Thousand Oaks: Sage.

Hajjem, A., Bellavance, F., & Larocque, D. (2014, June). Mixed-effects random forest for clustered data. *Journal of Statistical Computation and Simulation*, *84*(6), 1313–1328. doi: 10.1080/00949655.2012.741599

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction* (Second ed.). New York: Springer.

Hedges, L. V. (1981). Distribution Theory for Glass's Estimator of Effect Size and Related Estimators. *Journal of Educational Statistics*, *6*(2), 107–28.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and Random-effects Models in Meta-analysis. *Psychological Methods*, *3*(4), 486–504.

Higgins, J. P. T., & Thompson, S. G. (2002, June). Quantifying heterogeneity in a meta-analysis. *Statistics in Medicine*, *21*(11), 1539–1558. doi: 10.1002/sim.1186

Higgins, J. P. T., & Thompson, S. G. (2004). Controlling the risk of spurious findings from meta-regression. *Statistics in Medicine*, *23*(11), 1663–1682. doi: 10.1002/sim.1752

Higgins, J. P. T., Thompson, S. G., & Spiegelhalter, D. J. (2009, January). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society. Series A, (Statistics in Society)*, *172*(1), 137–159. doi: 10.1111/j.1467-985X.2008.00552.x

Janitza, S., Celik, E., & Boulesteix, A.-L. (2016, September). A computationally fast variable importance test for random forests for high-dimensional data. *Advances in Data Analysis and Classification*, 1–31. doi: 10.1007/s11634-016-0270-x

Karpievitch, Y. V., Hill, E. G., Leclerc, A. P., Dabney, A. R., & Almeida, J. S. (2009, September). An Introspective Comparison of Random Forest-Based Classifiers for the Analysis of Cluster-Correlated Data by Way of RF++. *PLOS ONE*, *4*(9),

e7087. doi: 10.1371/journal.pone.0007087

Laws, K. R. (2016, June). Psychology, replication & beyond. *BMC Psychology*, *4*, 30. doi: 10.1186/s40359-016-0135-2

Li, X., Dusseldorp, E., & Meulman, J. J. (2017). Meta-CART: A tool to identify interactions between moderators in meta-analysis. *British Journal of Mathematical and Statistical Psychology*.

Lyon, A. (2013). Why are normal distributions normal? *The British Journal for the Philosophy of Science*, *65*(3), 621–649. doi: 10.1093/bjps/axs046

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498. doi: 10.1037/a0039400

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011, February). Interpretation of random effects meta-analyses. *BMJ*, *342*, d549. doi: 10.1136/bmj.d549

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011, November). False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychological Science*, *22*(11), 1359–1366. doi: 10.1177/0956797611417632

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), 323–348. doi: 10.1037/a0016973

Svetnik, V., Liaw, A., Tong, C., & Wang, T. (2004). Application of Breiman's Random Forest to Modeling Structure-Activity Relationships of Pharmaceutical Molecules. In F. Roli, J. Kittler, & Windeatt (Eds.), *Lecture Notes in Computer Science: Multiple Classifier systems.* (pp. 334–343). Berlin, Heidelberg: Springer.

Team, R. C. (2017). *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing.

Thompson, S. G., & Higgins, J. P. T. (2002, June). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine*, *21*(11),

1559–1573. doi: 10.1002/sim.1187

Van Erp, S., Verhagen, J., Grasman, R. P., & Wagenmakers, E.-J. (2017). Estimates of Between-Study Heterogeneity for 705 Meta-Analyses Reported in Psychological Bulletin From 1990–2013. *Journal of Open Psychology Data*, *5*(1).

Viechtbauer, W. (2007, January). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*(1), 37–52. doi: 10.1002/sim.2514

Viechtbauer, W., & others. (2010). Conducting meta-analyses in R with the metafor package. *J Stat Softw*, *36*(3), 1–48.

Wright, M. N., & Ziegler, A. (2015, August). Ranger: A Fast Implementation of Random Forests for High Dimensional Data in C++ and R. *arXiv:1508.04409 [stat]*.

Table 1

Study 1: Partial η^2 for the influence of design factors on R^2_{cv}

	MetaForest	Fixed-effects	Uniform	metaCART	Δ_{MF-FX}	Δ_{MF-UN}	Δ_{MF-CA}
β	0.78	0.78	0.77	0.57	0.01	0.01	0.06
τ^2	0.19	0.19	0.16	0.08	0.00	0.00	0.00
k	0.29	0.29	0.28	0.53	0.00	0.00	0.25
$\bar{n}$	0.34	0.33	0.33	0.15	0.00	0.00	0.01
M	0.18	0.18	0.16	0.00	0.00	0.00	0.08
Model	0.49	0.49	0.46	0.36	0.00	0.00	0.03
β:τ^2	0.02	0.02	0.02	0.01	0.00	0.00	0.00
β:k	0.11	0.11	0.09	0.38	0.00	0.00	0.21
β:$\bar{n}$	0.08	0.08	0.07	0.05	0.00	0.00	0.00
β:M	0.12	0.12	0.11	0.00	0.00	0.00	0.06
β:Model	0.20	0.20	0.18	0.16	0.00	0.00	0.01
τ^2:k	0.01	0.01	0.01	0.03	0.00	0.00	0.02
τ^2:$\bar{n}$	0.04	0.05	0.04	0.02	0.00	0.00	0.00
τ^2:M	0.01	0.01	0.01	0.00	0.00	0.00	0.01
τ^2:Model	0.01	0.01	0.01	0.00	0.00	0.00	0.00
k:$\bar{n}$	0.02	0.02	0.01	0.08	0.00	0.00	0.03
k:M	0.00	0.00	0.00	0.00	0.00	0.00	0.00
k:Model	0.01	0.01	0.01	0.24	0.00	0.00	0.23
$\bar{n}$:M	0.02	0.02	0.02	0.00	0.00	0.00	0.01
$\bar{n}$:Model	0.02	0.02	0.02	0.01	0.00	0.00	0.00
M:Model	0.02	0.02	0.02	0.00	0.00	0.00	0.01

Note. Abbreviations: Random-effects MetaForest (MF), fixed-effects (FX), Unweighted (UN), and metaCART (CA).

Table 2

Study 2: Partial η^2 for the influence of design factors on R^2_{cv}

	MetaForest	Fixed-effects	Uniform	metaCART	Δ_{MF-UN}	Δ_{MF-FX}	Δ_{MF-CA}
β	0.50	0.44	0.48	0.21	0.13	0.00	0.05
τ^2	0.27	0.29	0.25	0.08	0.00	0.00	0.02
k	0.47	0.46	0.46	0.51	0.04	0.00	0.09
$\bar{n}$	0.06	0.06	0.05	0.01	0.00	0.00	0.01
M	0.10	0.12	0.09	0.00	0.00	0.00	0.05
Model	0.40	0.29	0.40	0.47	0.36	0.01	0.20
β:τ^2	0.06	0.06	0.05	0.01	0.00	0.00	0.01
β:k	0.03	0.02	0.03	0.10	0.01	0.00	0.04
β:$\bar{n}$	0.01	0.01	0.01	0.00	0.00	0.00	0.00
β:M	0.01	0.01	0.01	0.00	0.00	0.00	0.00
β:Model	0.26	0.38	0.24	0.14	0.12	0.00	0.01
τ^2:k	0.01	0.01	0.01	0.03	0.00	0.00	0.01
τ^2:$\bar{n}$	0.01	0.01	0.01	0.00	0.00	0.00	0.00
τ^2:M	0.00	0.00	0.00	0.00	0.00	0.00	0.00
τ^2:Model	0.10	0.12	0.09	0.05	0.00	0.00	0.00
k:$\bar{n}$	0.00	0.00	0.00	0.01	0.00	0.00	0.00
k:M	0.00	0.00	0.00	0.00	0.00	0.00	0.00
k:Model	0.04	0.03	0.05	0.26	0.03	0.01	0.25
$\bar{n}$:M	0.00	0.00	0.00	0.00	0.00	0.00	0.00
$\bar{n}$:Model	0.02	0.02	0.02	0.01	0.00	0.00	0.00
M:Model	0.04	0.04	0.04	0.00	0.00	0.00	0.02

Note. Abbreviations: Random-effects MetaForest (MF), fixed-effects (FX), Unweighted (UN), and metaCART (CA).

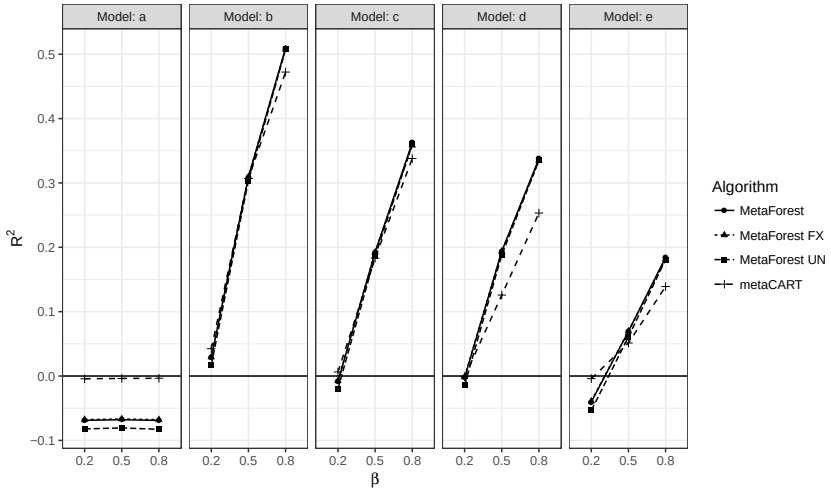

Figure 1. Study 1: Marginal R^2_{cv} for the interaction between effect size and model.

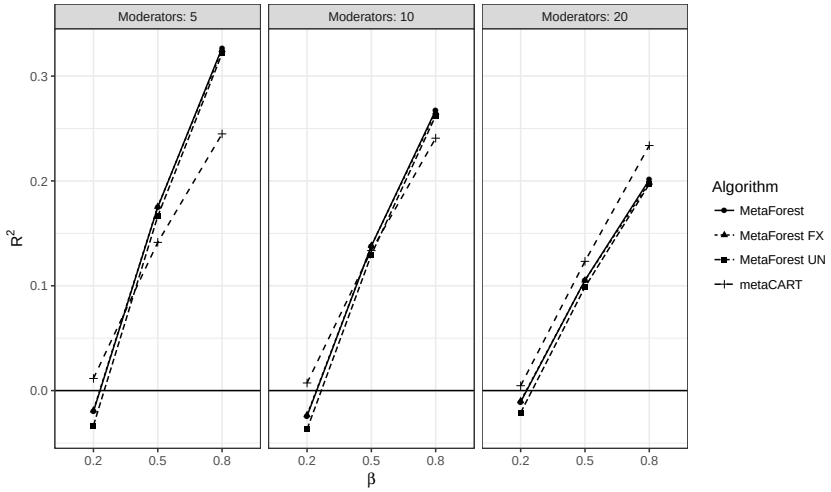

Figure 2. Study 1: Marginal R^2_{cv} for the interaction between effect size and the number of moderators.

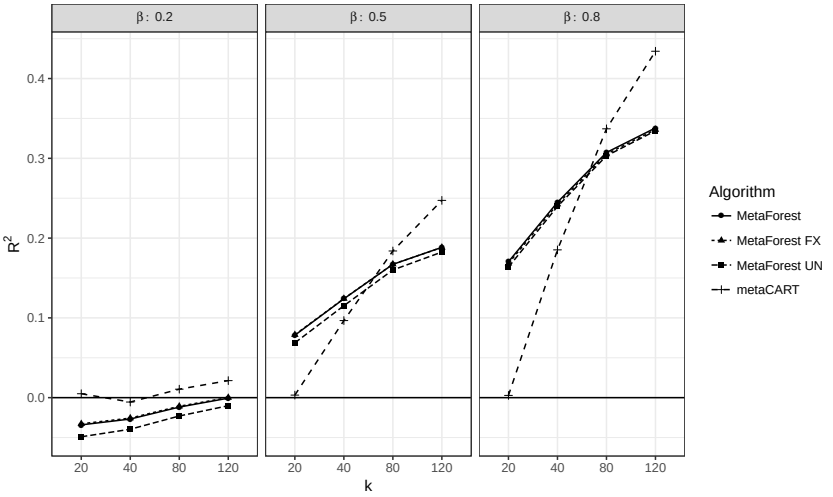

Figure 3. Study 1: Marginal R^2_{cv} for the interaction between effect size and the number of studies.

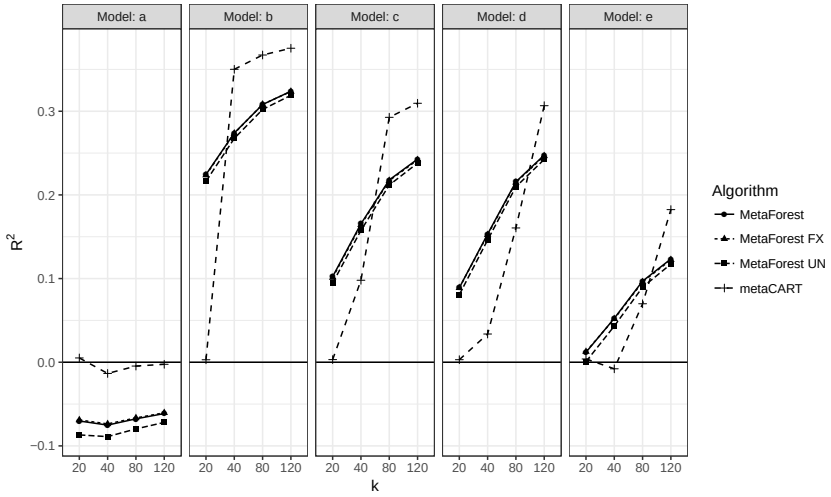

Figure 4. Study 1: Marginal R^2_{cv} for the interaction the number of studies and the true model.

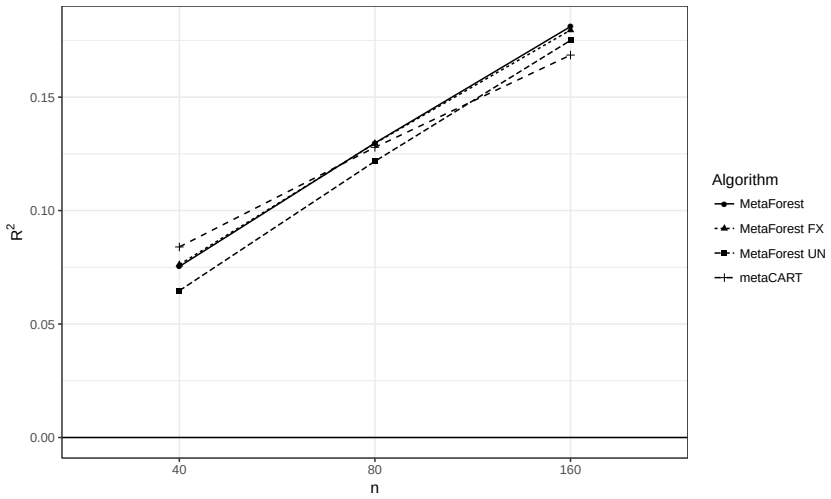

Figure 5. Study 1: Marginal R^2_{cv} for the effect of $\bar{n}$.

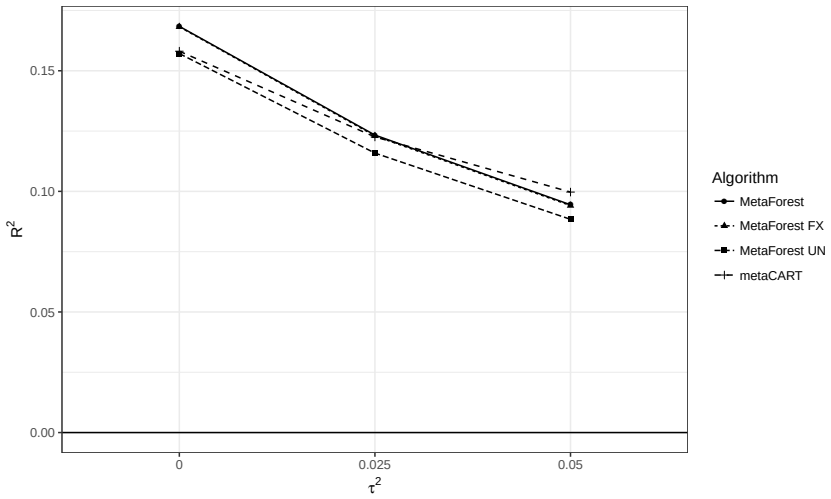

Figure 6. Study 1: Marginal R^2_{cv} for the effect of τ^2.

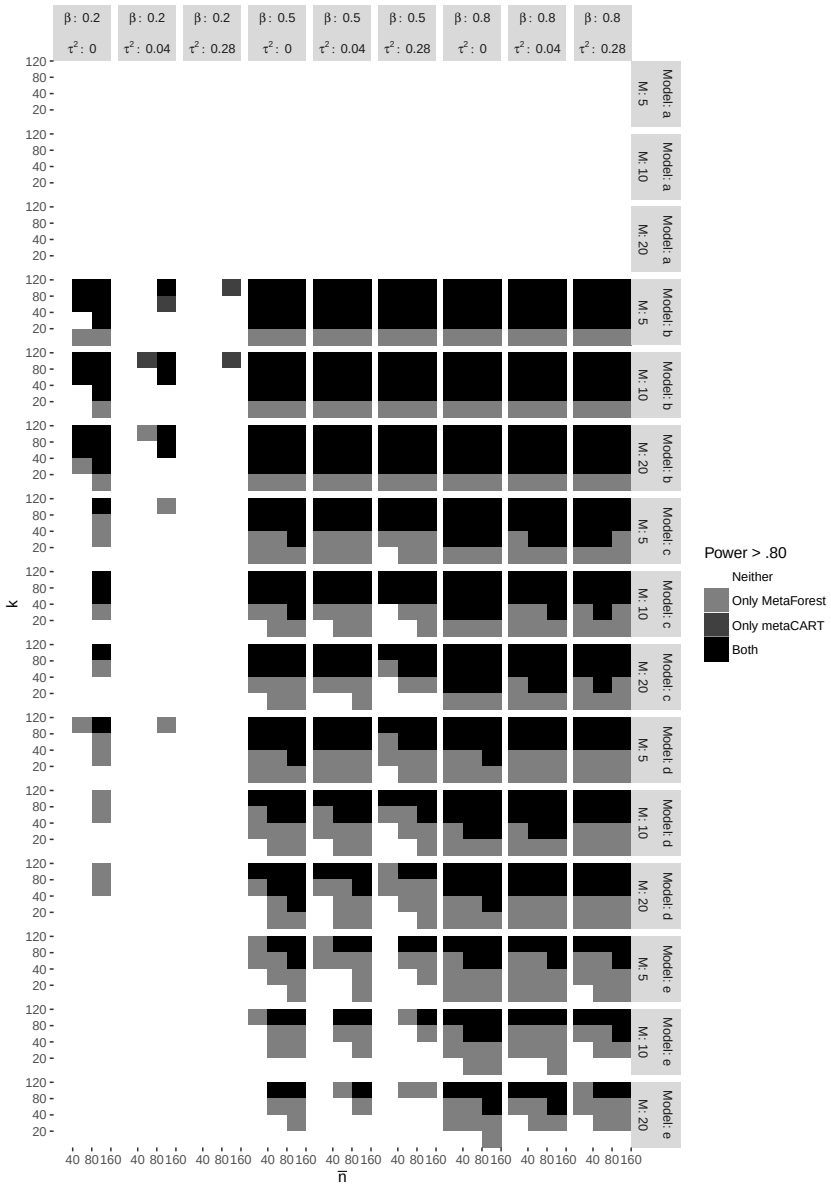

Figure 7. Study 1: Conditions under which random-effects MetaForest and metaCART reach 80% power.

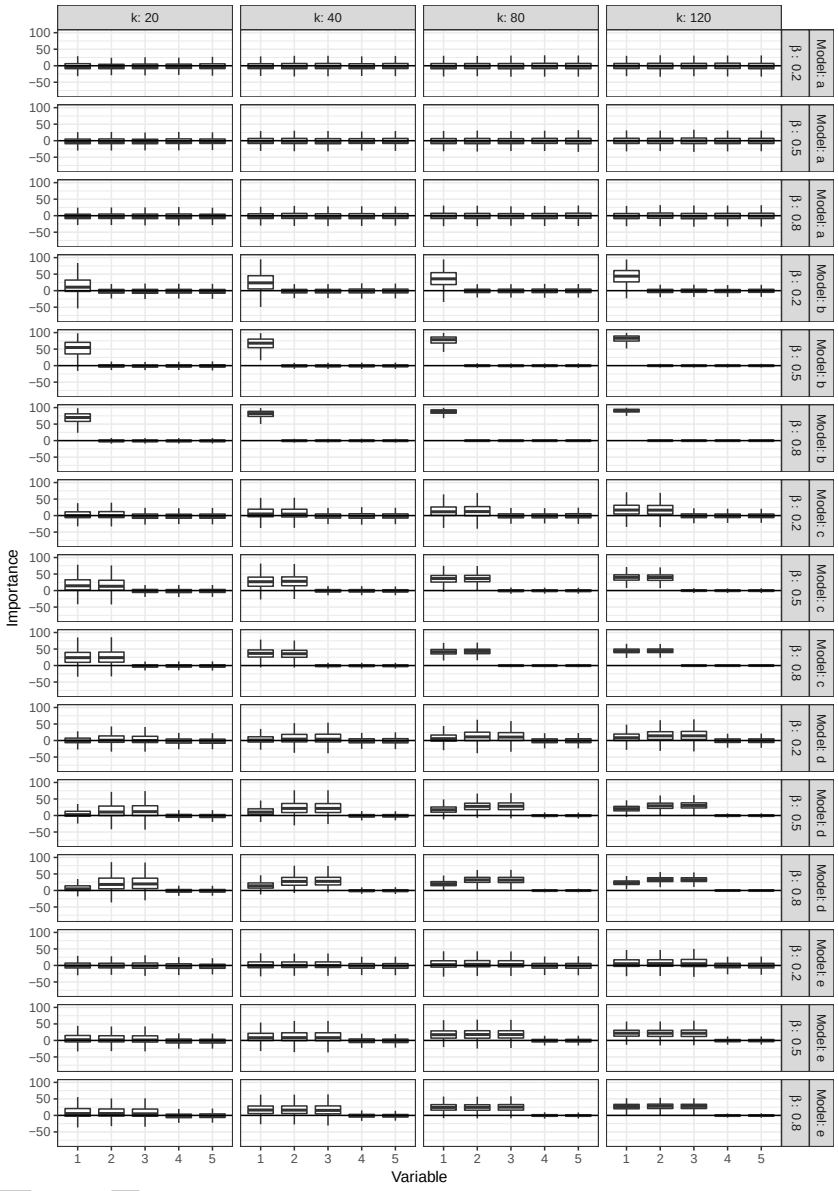

Figure 8. Study 1: Boxplots indicate the proportion of cases in which MetaForest assigned positive variable importance to a variable.

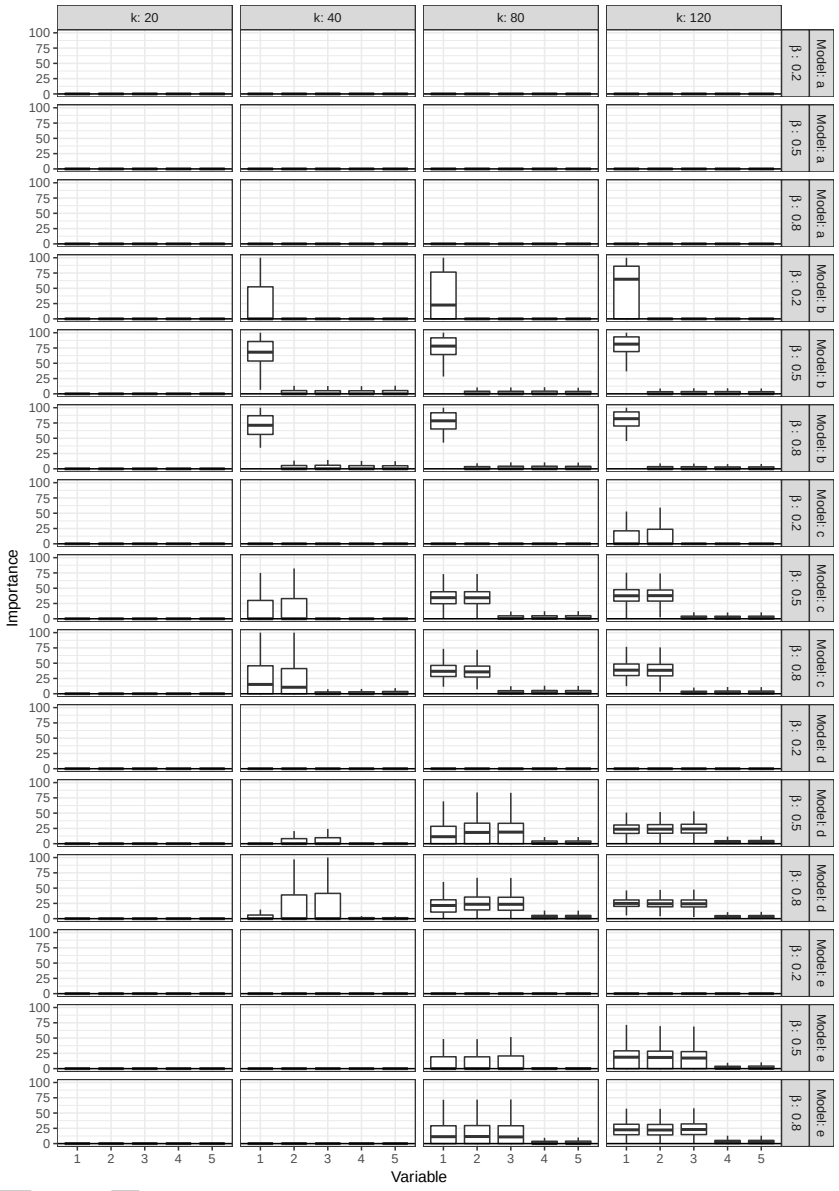

Figure 9. Study 1: Boxplots indicate the proportion of cases in which metaCART assigned positive variable importance to a variable.

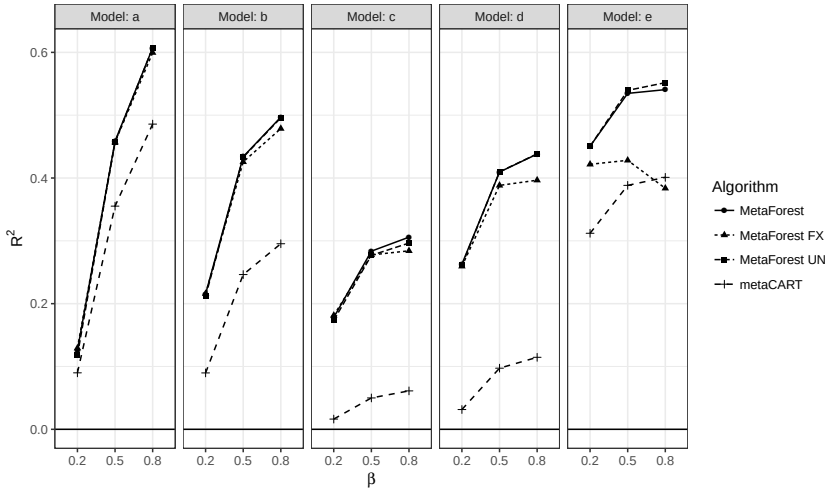

Figure 10. Study 2: Marginal R^2_{cv} for the interaction between effect size and the model.

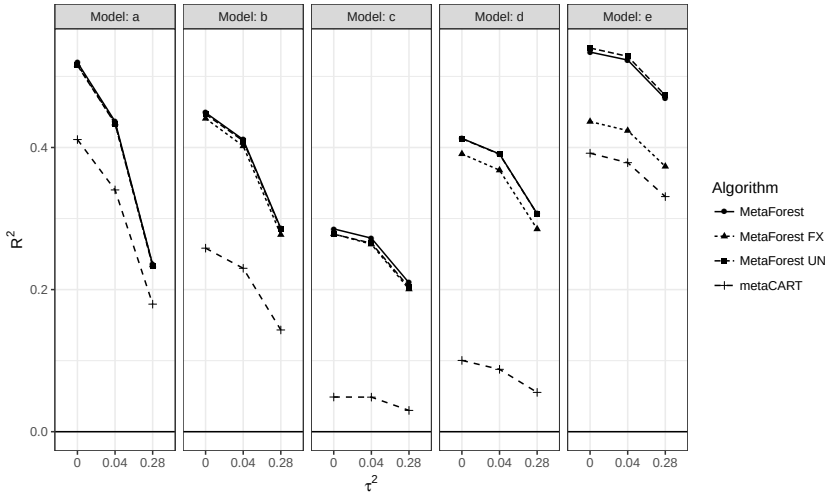

Figure 11. Study 2: Marginal R^2_{cv} for the interaction between residual heterogeneity and the model.

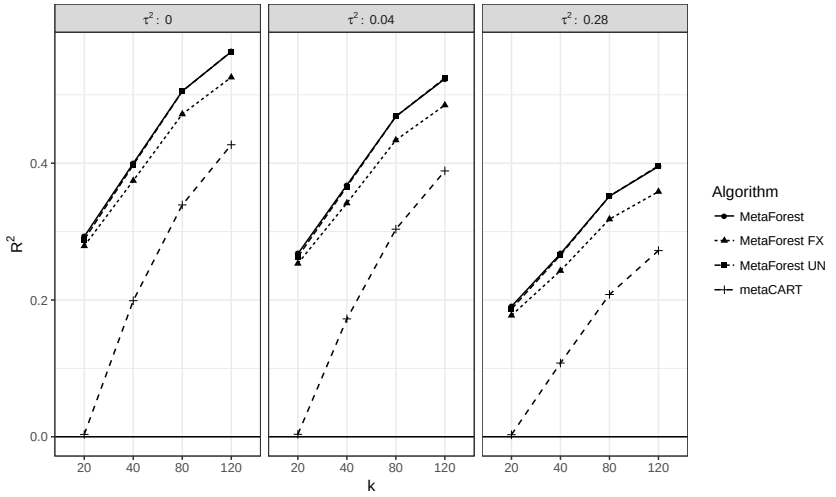

Figure 12. Study 2: Marginal R^2_{cv} for the interaction between the number of studies and residual heterogeneity.

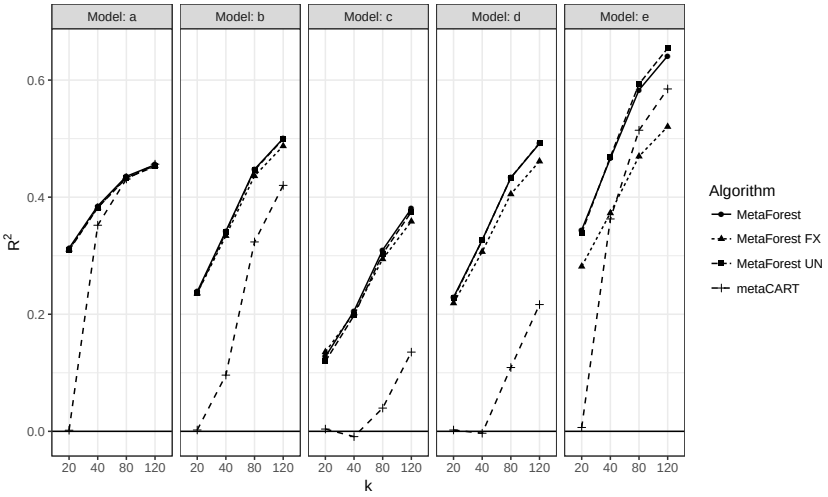

Figure 13. Study 2: Marginal R^2_{cv} for the interaction between the number of studies and the true model.

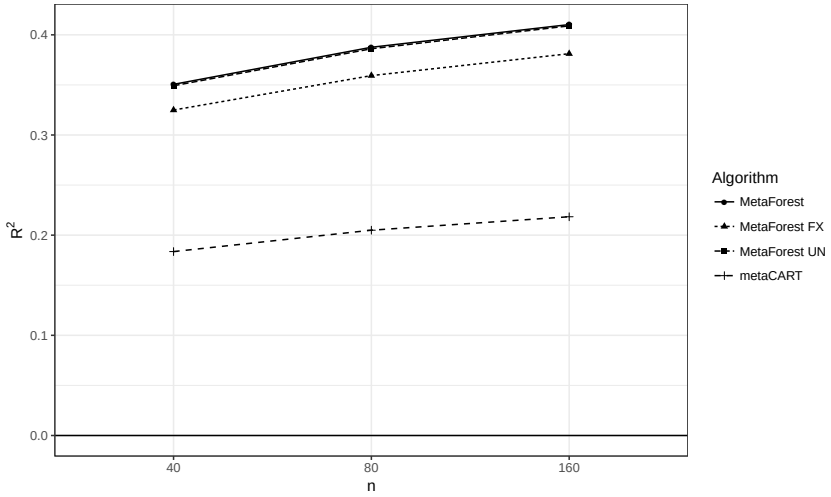

Figure 14. Study 2: Marginal R^2_{cv} for the effect of the average within-study sample size.

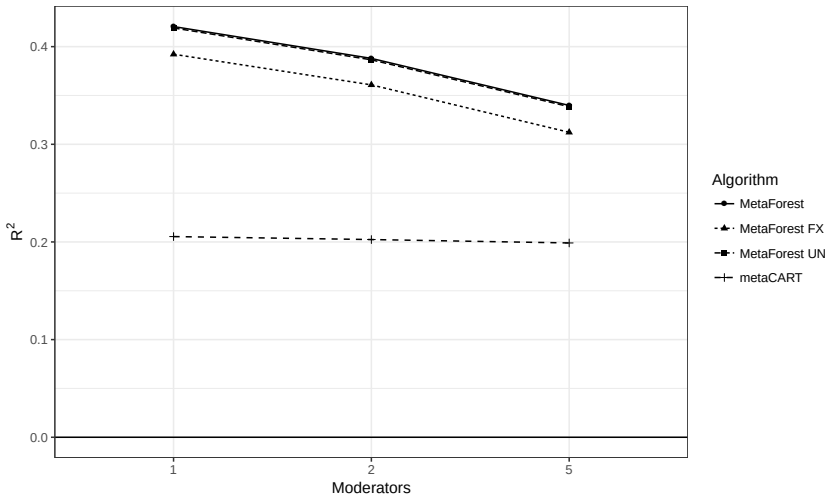

Figure 15. Study 2: Marginal R^2_{cv} for the effect of the number of irrelevant moderators.

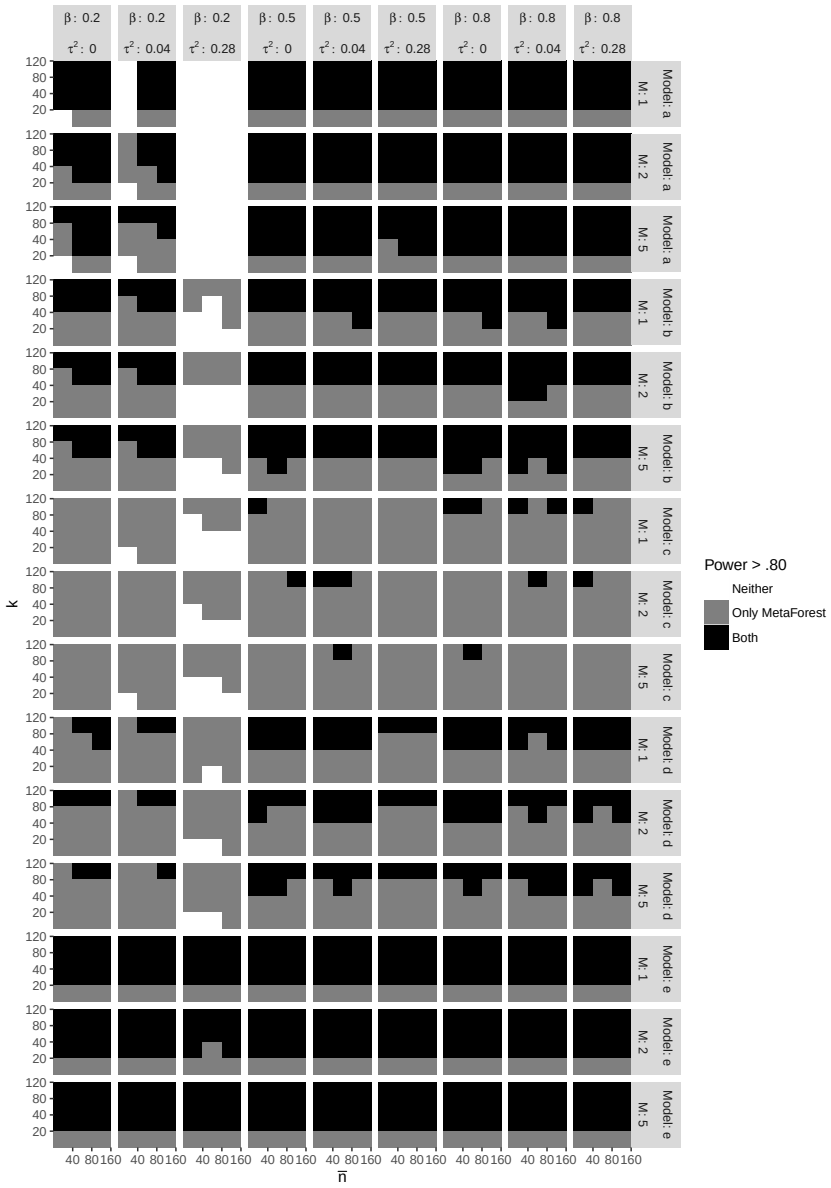

Figure 16. Study 2: Conditions under which random-effects MetaForest and metaCART reach 80% power.

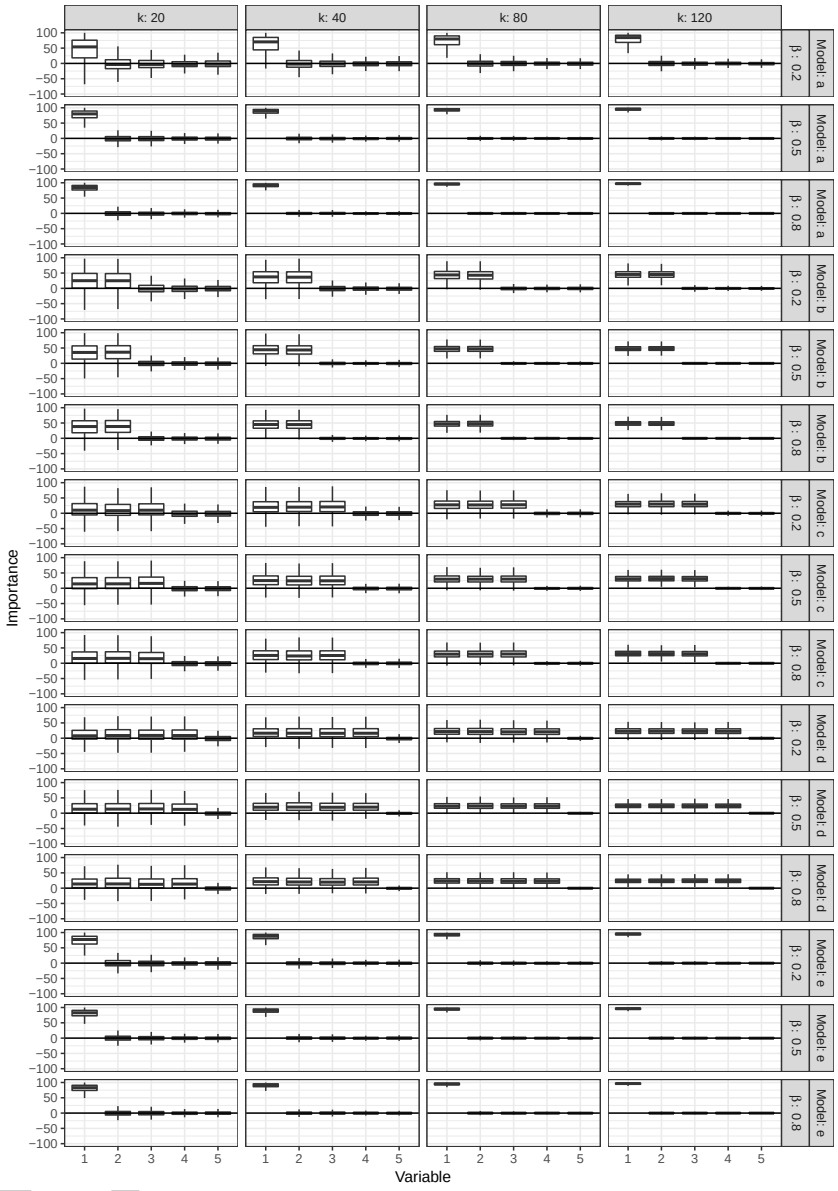

Figure 17. Study 2: Boxplots indicate the proportion of cases in which MetaForest assigned positive variable importance to a variable.

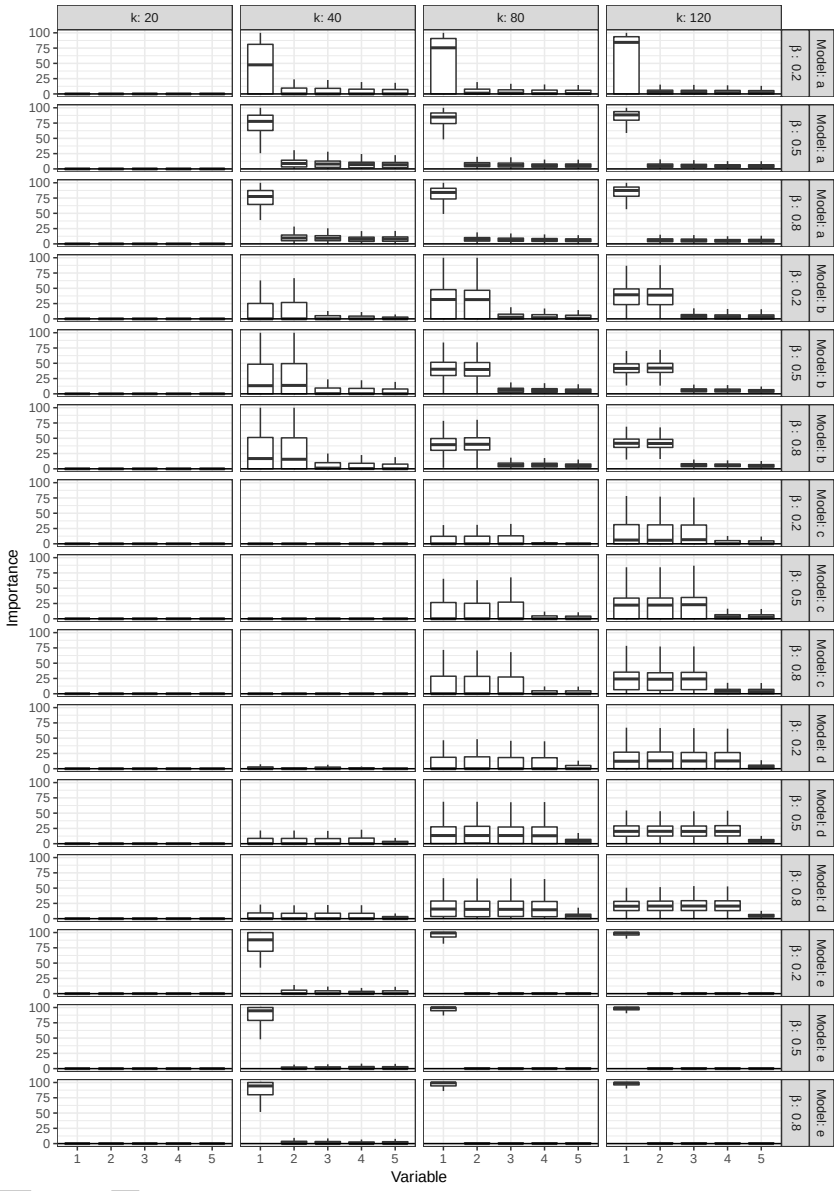

Figure 18. Study 2: Boxplots indicate the proportion of cases in which metaCART assigned positive variable importance to a variable.