The integration of meta-analysis and classification & regression trees: meta-CART

Li, X.

The integration of meta-analysis and classification & regression trees: meta-CART

Proefschrift
ter verkrijging van
de graad van Doctor aan de Universiteit Leiden
op gezag van Rector Magnificus Prof.mr. C. J. J. M. Stolker,
volgens besluit van het College voor Promoties
te verdedigen op Donderdag 27 Februari 2020
klokke 16:15 uur

door

Xinru Li
geboren te Baiyin, China
in 1988

Promotor:

Prof. dr. Jacqueline J. Meulman (Universiteit Leiden & Stanford University)

Copromotor:

Dr. Elise Dusseldorp (Universiteit Leiden)

Samenstelling van de promotiecommissie:

Prof. dr. Bart de Smit (Universiteit Leiden, voorzitter)

Prof. dr. Aad J.W. van der Vaart (Universiteit Leiden, secretaris)

Prof. dr. Marcel A.L.M. van Assen (Tilburg University)

Dr. Wolfgang Viechtbauer (Maastricht University)

Prof. dr. Xiaogang Su (The University of Texas at El Paso)

Contents

Chapter 1

Introduction

In recent years, the number of scientific publications has been rapidly growing. The same research question is often addressed by multiple independent studies, and tools are needed to synthesize and summarize the research findings on the same topic. Meta-analysis has been an increasingly popular and valuable tool to evaluate the evidence in areas as diverse as social and behavioral sciences, medicine, biology, economics and marketing, among others. The primary goal of a meta-analysis is to estimate a quantitative indicator of the strength of evidence (i.e., effect size) and the consistency of the evidence. To synthesize the research findings from multiple studies, meta-analysis combines the effect sizes computed in each separate study, and computes a weighted mean of the study effect sizes as a representative (called an overall effect size) for all studies. In practice, heterogeneity often exists in the study effect sizes due to the variations, for example, in the sample characteristics, the implementations of the experiments or treatments, the measurement points, and other factors that may influence the study outcomes. Heterogeneity may lead to contradictory conclusions, and makes it misleading to summarize all the study results by only one overall effect size. In such cases, searching for the factors that account for the heterogeneity in study effect sizes can be of high interest. So-called moderator analysis is used to assess the contribution of the study-level factors on the heterogeneity.

Over the past decades, moderator analysis has focused on using regression, called "meta-regression", to understand the relationship between the effect size and moderators. However, when multiple potential moderators are available, meta-regression has several limitations. First, in the case of a large number of potential moderators, meta-regression analysis has limited power to simultaneously investigate their influence. Second, meta-regression requires the potential moderators and their interaction effects to be specified beforehand, and usually it is difficult to include all possible interaction effects in one model. Thus, interactions between moderators are seldom investigated in meta-analytic studies. The knowledge of interaction effects, however, can provide valuable information to understand whether the moderators amplify or attenuate each other's influence on the effect size. Moreover, such knowledge can be essential if the study is aiming at identifying

the optimal intervention for specific subpopulations or the intervention with the most effective combination of treatment components.

To overcome the aforementioned limitations, we propose using tree-based methods in the framework of meta-analysis to explore the influence of moderators and examine the interaction effects between them. Tree-based methods were introduced for the first time by Morgan and Sonquist (1963) in a method called automatic interaction detection (AID), and fully developed in classification and regression trees (CART) by Breiman, Friedman, Stone, and Olshen (1984). Tree-based methods have the advantage of dealing with interactions and handling non-linear relationships, and produce analysis results that can be easily interpreted. In many fields, trees are popular approaches in single study settings. But the application of tree-based methods in the framework of meta-analysis has seldom been studied. In addition, to date, the field of meta-analysis has primarily focused on confirmatory analyses, and methods development in explanatory analyses aiming at explaining heterogeneity has been quite limited (Tipton, Pustejovsky, & Ahmadi, 2018).

Therefore, in this thesis a tree-based method, called meta-CART, is introduced for investigating moderator effects in meta-analysis and the interactions among them. The idea of meta-CART was originally proposed by Dusseldorp, van Genugten, van Buuren, Verheijden, and van Empelen (2014). The researchers apply a two-step procedure by first using classification and regression trees (CART) to identify interactions, and then performing subgroup meta-analysis to test the significance of moderator effects. From the original study, several issues arose. First, the study only considered one application, and no simulation study was performed. Second, it used a classification tree in the first step of the analysis by dichotomizing the study effect sizes by their median. Hence, it is interesting to investigate whether using a regression tree will have better performance than using a classification tree. Third, the studies were not weighted by their accuracies. This thesis addresses these issues by further developing the meta-CART method and performing extensive simulation studies to examine the performance of the proposed new implementations. An important aim of these simulation studies is to find the conditions under which meta-CART can achieve satisfactory performance. Further aims are to compare the performance of different implementations and to recommend guidelines for practice.

The outline of this thesis is as follows. In Chapter 2, new meta-CART extensions are proposed by using a regression tree to avoid dichotomization, by weighting study effect sizes with their accuracy, and applying different pruning rules. The performance of these extensions is evaluated via an extensive Monte Carlo simulation study on dichotomous moderator variables. The factors that may influence the performance of meta-CART are investigated.

In Chapter 3, new algorithms are proposed for meta-CART by integrating the two steps of the approach into one step and by consistently taking into account the fixed-

effect or random-effects assumption in both the tree-growing and the subgroup analysis processes. For fixed effect meta-CART, weights are applied, and the subgroup analysis is adapted. For random effects meta-CART, a new algorithm has been developed. The performance of this new version of meta-CART is investigated via an extensive simulation study on different types of moderator variables (i.e., dichotomous, nominal, ordinal, and continuous variables).

Chapter 4 introduces the R-package `metacart` that provides user-friendly functions to conduct meta-CART analyses in **R**. The core features of `metacart` are described, and the application of this package is illustrated with diverse examples.

In Chapter 5, meta-CART is applied to a real-world data set to explore the interactions among moderators. This meta-analytic study explores the influence of behavior change techniques (BCTs) and the context or delivery components on the effectiveness of healthy eating, physical activity and smoking interventions for low-income groups.

In Chapter 6, the meta-CART method of Chapter 3 is further extended in four aspects: for the tree-growing process, a smooth sigmoid surrogate (SSS) strategy and a look-ahead strategy are proposed to alleviate the local optimum problem and speed up the splitting procedure. For the statistical inference in the identified subgroups, a permutation test is used to appropriately assess the statistical significance of the heterogeneity between identified subgroups. Finally, a bootstrap procedure is proposed to correct the over-optimism in the confidence intervals of the estimated subgroup effect sizes. The performance of the proposed new approaches is evaluated via a simulation study. And the application is illustrated with a real-world data.

In the epilogue, the main results presented in this thesis are summarized. Also, the limitations of my work are discussed and suggestions for future research are made.

Publications

Chapters 2-6 are based on published or submitted papers. Hence, they can be read separately. Here follows a list of the papers.

- Chapter 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*, 70(1), 118-136.

- Chapter 3: Li, X., Dusseldorp, E., & Meulman, J. J. (2019). A flexible approach to identify interaction effects between moderators in meta-analysis. *Research Synthesis Methods*, 10(1), 134-152.

- Chapter 4: Li, X., Dusseldorp, E., Su, X., & Meulman, J.J. (in press), "Multivariate moderator meta-analysis with the R-package metacart". *Behavior Research Methods*

- Chapter 5: Bull, E. R., McCleary, N., Li, X., Dombrowski, S. U., Dusseldorp, E., & Johnston, M. (2018). Interventions to promote healthy eating, physical activity and smoking in low-income groups: a systematic review with meta-analysis of behavior change techniques and delivery/context. *International Journal of Behavioral Medicine*, 1-12.

- Chapter 6: Li, X., Dusseldorp, E., Su, X., & Meulman, J.J. (submitted), Advanced tree-based subgroup identification in meta-analysis.

Chapter 2

Meta-CART: A tool to identify interactions between moderators in meta-analysis

abstract

In the framework of meta-analysis, moderator analysis is usually performed only univariately. When several study characteristics are available that may account for treatment effect, standard meta-regression has difficulties in identifying interactions between them. To overcome this problem, meta-CART has been proposed: an approach that performs Classification and Regression Trees (CART) to identify interactions, and then subgroup meta-analysis to test the significance of moderator effects. The previous version of meta-CART has its shortcomings: when applying CART, the sample sizes of studies are not taken into account, and the effect sizes are dichotomized around the median value. Therefore, this article proposes new meta-CART extensions, with weighting study effect sizes by their accuracy, and with performing a regression tree to avoid dichotomization. In addition, new pruning rules are proposed. The performance of all versions of meta-CART was evaluated via a Monte Carlo simulation study. The simulation results revealed that meta-regression trees with random-effects weights and half a standard-error pruning rule have the best performance. The required sample size for meta-CART to achieve satisfactory performance depends on the number of study characteristics, the magnitude of the interactions, and the residual heterogeneity.

2.1 Introduction

In psychology and medicine, meta-analysis is a powerful tool to quantitatively integrate findings from multiple studies in a systematic way. The effect size is usually chosen as the standard representative of the study results. By combining the effect sizes, meta-analysis computes a weighted mean of the study effect sizes as a representative (called a summary effect size) for all studies. However, in case of substantial heterogeneity between the studies, the summary effect size is not a good representative for all studies. In such cases, it is important to identify possible causes of the heterogeneity (Normand, 1999; Thompson, 1994; Thompson & Sharp, 1999). The search for study characteristics (e.g., quality of the design) that might account for effect size heterogeneity is called moderator analysis and such study characteristics are called "moderators". The most popular types of moderator analyses are subgroup meta-analysis (for a categorical moderator) and meta-regression (for a continuous one).

In most meta-analytic studies only univariate moderator analysis is performed (e.g., Huisman, De Gucht, Dusseldorp, & Maes, 2009). A plausible reason for this is that popular meta-analysis programs do not allow to include multiple moderators into one analysis (e.g., Comprehensive Meta-Analysis; Borenstein, Hedges, Higgins, & Rothstein, 2009). However, some recently developed programs allow for meta-regression with multiple moderators (e.g., the R-package metafor; Viechtbauer, 2010).

Besides the need of including multiple moderator variables into one analysis, recent meta-analyses emphasize the need to model interaction effects between moderators. When interventions consist of several components, the researcher might be interested not only in a research question such as "Are the interventions generally effective?" but also in "Which combinations of components have the greatest probability of being most effective?" (Welton, Caldwell, Adamopoulos, & Vedhara, 2009). For example, in a meta-analysis on the influence of behaviour change techniques (BCTs) on the effect of physical activity and healthy eating interventions, the authors discussed that the data strongly suggest that inclusion of a specific BCT (i.e., self-monitoring) in combination with other BCTs (e.g., self-regulation techniques) is likely to enhance the effectiveness of interventions. (Michie, Abraham, Whittington, McAteer, & Gupta, 2009). When a priori hypotheses exist, standard meta-regression can be used to investigate interaction effects. However, interaction effects between moderators are seldom investigated in meta-analysis. One possible reason is the lack of theory and previous findings on possible interaction effects. When no a priori hypotheses exist, the nature of the study is usually exploratory. In such cases, standard meta-regression often lacks enough power for interaction detection between multiple moderators. Furthermore, it is difficult for standard meta-regression to investigate higher-order interaction effects. For example, in a meta-analysis of 50 studies with 10 study characteristics that might account for the heterogeneity (i.e., potential modera-

tors), there are $\binom{10}{2}$, that is, 45, possible two-way interaction terms and $\binom{10}{3}$, that is, 120, possible three-way interaction terms. In such cases, it is not possible for a standard meta-regression to include all the interaction terms simultaneously.

Recently, a new approach called meta-CART was proposed to overcome these difficulties (Dusseldorp et al., 2014). In a situation with many available study characteristics, meta-CART searches for those combinations of characteristics that might account for effect size heterogeneity. The method is a combination of Classification and Regression Trees (CART; Breiman et al., 1984) and subgroup meta-analysis. In the first step of meta-CART, a tree is fitted by CART using the study effect sizes as response variable, and the study characteristics as predictor variables. In the second step, the terminal nodes of the tree are used to create a new subgroup variable (with categories referring to the labels of the leaves in which the studies were assigned to by CART) and a standard subgroup meta-analysis is performed using the new subgrouping variable as moderator. Initial results of meta-CART were promising from a substantial point of view (Dusseldorp et al., 2014), that is, the results could be easily interpreted and were meaningful. Also, the potential of the approach has been acknowledged (Michie, Johnson, & Johnston, 2015; O'Brien et al., 2015). However, the recovery performance of meta-CART has not been investigated yet. Furthermore, the previous version of meta-CART has several shortcomings: (a) it uses a classification tree in the first step of the analysis. To obtain a distinction between more successful and less successful interventions, the study effect sizes are dichotomized, which implies loss of information (Hunter & Schmidt, 1990); (b) The sample sizes of studies are not taken into account when applying CART, which means that CART ignores the accuracy of the effect-size estimates.

In this paper, the first goal is to address possible solutions to overcome these shortcomings, by omitting the dichotomization of the response variable, and by weighting study effect sizes by their estimate accuracy. In addition, new pruning rules are proposed to improve the performance of meta-CART. The proposed methodology results in two types of trees (i.e., meta-classification trees and meta-regression trees), and several options of weights and pruning rules for each type of tree. The second goal of this paper is to compare the performance of all the options for meta-classification trees and meta-regression trees, and, if possible, choose the best options for each type of tree. In addition, the conditions for each type of tree to achieve satisfactory performance are explored. In this paper, we focus on the interaction effects between dichotomous study characteristics (e.g., BCTs).The outline of this paper is as follows. First, we introduce meta-analysis, CART and meta-CART. Next, we describe the proposed extensions of meta-CART. We then evaluate and compare the performance of all the options for meta-classification trees and also for meta-regression trees in an extensive simulation study. Depending on the results of the extensive simulation study, the best options for each type of tree will be selected. Finally, we summarize and discuss the results.

2.2 Meta-analysis

The main purpose of meta-analysis can be summarized in three objectives: (a) to synthesize the results of the studies; (b) to assess the heterogeneity in the studies, and (c) to search for moderators that can explain the heterogeneity (Sánchez-Meca & Marín-Martínez, 1998). Depending on the type of studies, a variety of different effect size measures can be used for a meta-analysis, including odds ratio, relative risk, correlation coefficient, and (standardized) mean difference. In this paper, we focus on studies that compare treatment and control groups with respect to some continuous response variable, and Hedges' g (Hedges, 1981) is used as the measure of effect size. Several tests can be employed to determine whether heterogeneity exists in the effect sizes, of which the Q-test is the most frequently used test. The formulas concerning effect size and the Q-statistic can be found in Appendix A.

2.2.1 Meta-regression and subgroup meta-analysis

Meta-regression investigates whether particular study characteristics explain any of the heterogeneity between studies. It can be performed under the fixed effects or the random effects model. Fixed effects meta-regression assumes that the influential study characteristics (i.e., moderators) explain all the heterogeneity between studies. Denote the true effect size in the k^{th} study by δ_k, and denote the observed effect size in the k^{th} study by g_k. Under the fixed effects assumption, the observed effect size is given by

$$g_k = \delta_k + \epsilon_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_M x_{Mk} + \epsilon_k, \tag{2.1}$$

where $x_{mk}(m = 1, ..., M)$ specify the study characteristics of the k^{th} study, and the βs are the corresponding regression coefficients. The sampling error ϵ_k is assumed to be distributed as $\mathcal{N}(0, \sigma^2_{\epsilon_k})$, where $\sigma^2_{\epsilon_k}$ is the sampling variance.

Random effects meta-regression allows for heterogeneity unexplained by moderators. In a random effects model, there are two sources to account for the total variance of δ_k: the variability introduced by the moderators in the model, and the additional variability introduced by other unmeasured factors. Such additional variability is called "residual heterogeneity" (Viechtbauer, 2007a), which will be denoted by σ^2_τ. Under the random effects assumption g_k is given by

$$g_k = \delta_k + \epsilon_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_M x_{Mk} + \tau_k + \epsilon_k, \tag{2.2}$$

where τ_k is distributed as $\mathcal{N}(0, \sigma^2_\tau)$, and it reflects that the true effect size may vary from study to study.

Subgroup meta-analysis assesses the relationship between *subgroup* membership and effect size, and is similar to meta-regression with one categorical study characteristic. In subgroup meta-analysis, we consider two types of models, depending on the within-subgroup and between-subgroups assumptions. The first one is the fixed effects model, which assumes fixed effects within subgroups and also across subgroups. This means that the difference in effect sizes between subgroups can be explained by the subgroup membership, and within each subgroup the studies share a common effect size. The second one is the mixed effects model, which is generally advocated in subgroup meta-analysis (Borenstein et al., 2009; Viechtbauer, 2010). The mixed effects model assumes a random effects model within subgroups and a fixed effects model across subgroups. In such a model, the difference in effect sizes between subgroups are all explained by the subgroup membership, and within each subgroup the model allows heterogeneity between studies. The details concerning the heterogeneity test in mixed effects models can be found in Appendix A.2.

2.3 CART

2.3.1 Building the tree

CART is a recursive partitioning method that was proposed by Breiman, Friedman, Stone, and Olshen (1984). The method can be used for modeling the relationships between predictor variables and a categorical response variable by a $classification$ tree or a continuous response variable by a $regression$ tree. For classification trees, trees are built by finding the split on a predictor variable that best discriminates between different classes of the response variable. This "best discriminates" is defined in terms of a partitioning criterion called the impurity function. For regression trees, the partitioning criterion is defined as the split that minimizes the squared difference between the observed and predicted values of the response variable. For a useful introduction into CART we refer to Merkle and Shaffer (2011) and more details concerning the partitioning criteria can be found in Appendix A.3.

2.3.2 Pruning the tree

To prevent overfitting, a recommended strategy is to first grow an initial tree by continuing the splitting process until all terminal nodes are either small (e.g., containing only one or two subjects, that is, one or two studies in our case) or with zero impurity. Then the initial tree is reduced to a final tree of smaller size by "pruning" the non-influential splits (Breiman et al., 1984).

In most applications, cross-validation is the preferred method to estimate the misclassification rate or sum-of-squared error. In this way, overfitting can be prevented and the

best size of the tree can be selected. Sometimes, the minimum cross-validation rule is used, by which the tree with the minimum cross-validation error is selected as the final tree. But the minimum cross-validation error may be unstable due to the uncertainty of its estimate. Therefore, Breiman et al. (1984) suggested using the one-standard-error rule to reduce the instability, which selects the smallest tree whose cross-validation error is within the minimum cross-validation error plus one standard error. To generalize the pruning rules, a pruning parameter c can be introduced to select the pruned tree by using the c·SE rule (Dusseldorp, Conversano, & Van Os, 2010). The c·SE rule selects the smallest tree whose cross-validation error is within the minimum cross-validation error plus the standard error multiplied by c. The one-standard-error rule and the minimum cross-validation error rule can be regarded as special cases of the c·SE rule when $c = 1$ and $c = 0$, respectively.

2.4 Meta-CART

2.4.1 Previous version of meta-CART

The previous version of meta-CART as proposed in Dusseldorp et al. (2014) is a two-step procedure. In the first step, a classification tree is fitted to detect interaction effects between multiple moderators using the dichotomized effect sizes as response variable. For an example of such a tree, we show a result from Dusseldorp et al. (2014) (see Figure 2.1). The aim of this study was to identify particular combinations of behaviour change techniques (BCTs) that explain intervention success (defined as an effect size higher than the overall effect size of all studies). The tree in Figure 2.1 represents an interaction between the BCTs "prompt intention formation" and "provide information about behaviour-health link". According to the classification tree, when the two BCTs are both included in the intervention, the percentage of more successful interventions is higher (77%) than interventions that include only one or none of the two BCTs (41% and 36%, respectively).
In the second step, a standard subgroup meta-analysis is performed to investigate whether the subgroup membership obtained from the pruned tree accounts for the heterogeneity between the studies. If the pruned tree obtained by the first step has two or more terminal

Table 2.1: Results of Subgroup Analysis Using a Mixed Effects Model

Group	# interv.	$\bar{g}$	95% CI	$Q(df)$	p value
Grouping variable of tree					
Group 1	33	0.26	0.16,0.35		
Group 2	51	0.24	0.18,0.29		
Group 3	22	0.46	0.39,0.59		
				25.2(2)	<0.001

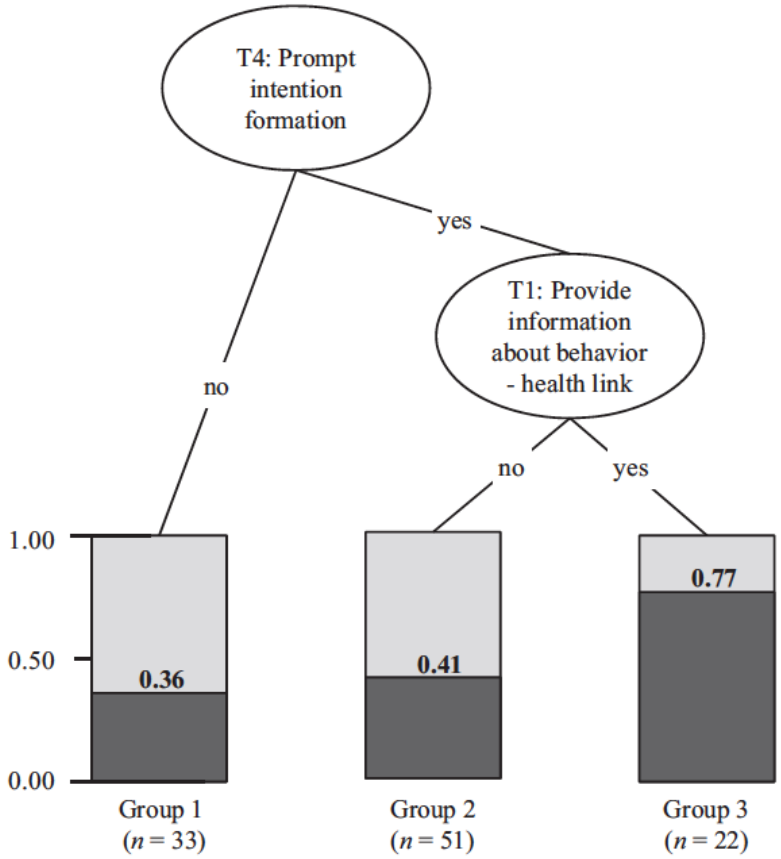


Figure 2.1: Classification tree for the studies that use at least one of the motivation-enhancing techniques from this category. Plots in the end nodes display the percentage of interventions that were more successful (i.e., an effect size higher than 0.31). T1 and T4 refer to Techniques 1 and 4 from the motivation-enhancing techniques category.

nodes, and the subgroup meta-analysis indicates significance between subgroups heterogeneity, meta-CART indicates the presence of (a) moderator effect(s). For example, the tree in Figure 2.1 results in three subgroups defined by the terminal nodes of the tree (rectangles). Consequently, one new variable was added, with three categories referring to these three terminal nodes. A subgroup meta-analysis was employed to test the significance of the new subgroup variable (see Table 2.1). As a result, the between subgroups Q-statistic indicates that the difference between these subgroups in mean effect sizes was highly significant. Interventions that included both "Prompt intention formation" and "Provide information" about behaviour-health link were, on average, more effective ($\bar{g} = 0.46$) than the other two groups of interventions ($\bar{g} = 0.24$ or $\bar{g} = 0.26$).

2.4.2 The extensions of meta-CART

As mentioned before, the previous version of meta-CART has two shortcomings: (a) the dichotomization of the effect sizes, and (b) the ignorance of the difference in accuracy of the effect sizes. In addition, the previous version uses the "one-standard-error" rule (see section 2.3.2) to prune the classification tree, and the performance of the pruning rule in meta-CART has not been investigated yet.

Regarding these shortcomings, we propose the following extensions that might improve the performance of meta-CART. First, the original numeric values of effect sizes can be used as the response variable to fit a regression tree, instead of fitting a classification tree. As a result, there can be two types of trees: meta-regression trees and meta-classification trees.

Second, weights can be assigned to the studies when applying CART. There are several possible types of weights depending on the assumption for the residual heterogeneity (the fixed effects or random effects model). Under the fixed effects assumption, the weight for each study can be computed as

$$w_k = \frac{1}{\hat{\sigma}^2_{\epsilon_k}} \Big/ \sum_{k=1}^{K} \frac{1}{\hat{\sigma}^2_{\epsilon_k}}. \tag{2.3}$$

The fixed effects weights do not take into account the residual heterogeneity. The weights under the random effects assumption are given by: (Cohen, 1988)

$$w_k = \frac{1}{\hat{\sigma}^2_{\epsilon_k} + \hat{\sigma}^2_{\tau}} \Big/ \sum_{k=1}^{K} \frac{1}{\hat{\sigma}^2_{\epsilon_k} + \hat{\sigma}^2_{\tau}}. \tag{2.4}$$

In total, there can be three types of weights for meta-CART: (a) all weights equal to 1, which is equivalent to applying no weights, (b) fixed effects weights, and (c) random effects weights. For convenience, these three types of weights will be denoted by $\mathcal{W}_0$, $\mathcal{W}_1$

and $\mathcal{W}_2$, respectively.

Third, the c·SE pruning rules (see section 2.3.2) can be employed in the partitioning procedure of meta-CART. The pruning rule may influence the detection rate of interaction effects. A pruning rule with a small value of c might be too liberal: the pruned tree obtained by CART appears to be too large, implying too many interaction terms. On the other hand, a pruning rule with a large value of c might be too conservative: the pruned tree is too small, resulting in too few or even no interaction terms. In order to find the optimal pruning rule for meta-CART, three values of c were chosen as 0, 0.5, 1.0.

Considering all the options, there are three options for the weights, and three options for the pruning rules. As a result, $3 \times 3 = 9$ possible options were proposed for each type of tree (meta-regression tree and meta-classification tree).

2.5 Simulation

2.5.1 Motivation

In the simulation study, we were interested in two questions: (a) which options of meta-CART generally have the best performance for each type of tree, and (b) given the best options, which conditions are influencing the performance of each type of tree? The conditions included observable features of meta-analytic data sets, such as the number of studies, the within-study sample sizes, and the number of study characteristics, as well as unobservable structures and parameters underlying the data, such as the complexity of the interaction effects, the magnitude of the interaction effect, and the residual heterogeneity.

The recovery performance of meta-CART was measured by the ability of successfully retrieving the true models underlying the data. Five tree structures were designed with increasing complexity as the underlying true model to generate data sets (see Figure 2.2). Model A was created to assess the probability that meta-CART falsely detects (a) moderator effect(s) when there is no moderator in the true model (Type I error). Model B was created to evaluate the ability of meta-CART to detect the main effect of a single moderator. Models C, D and E were created to evaluate the ability of meta-CART to correctly detect the interaction effects between moderators when interaction effects are present in the true model. In models C, D and E, the interventions are effective only in studies with certain combination(s) of study characteristics. For example, in model C the interventions are effective only when study characteristics x_1 and x_2 are both present. The studies are thereby split on moderators into subgroups. The average effect size in the ineffective subgroups was fixed to be 0. The average effect size in the effective subgroups was made a design factor and denoted by δ_I.

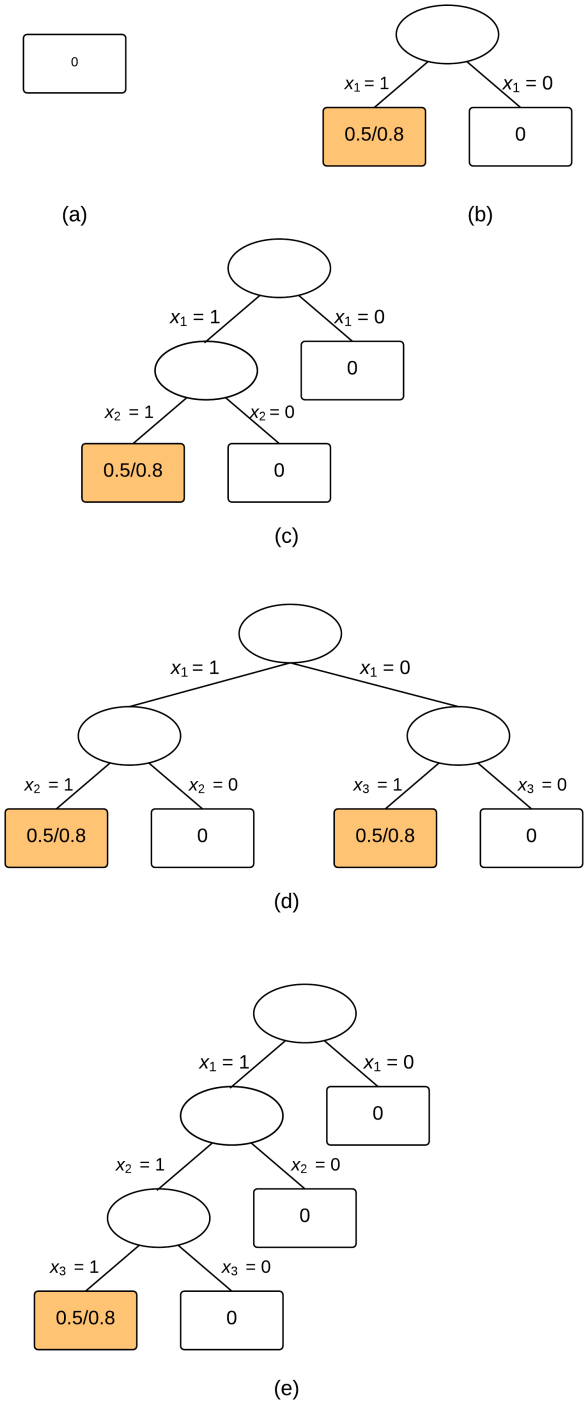


Figure 2.2: Simulated data sets were generated from five true tree structures: (a) to (e). These tree structures represents a true model including: no moderator effect (model A); only main effect of one moderator (model B); one two-way interaction (model C); two two-way interactions (model D); and one three-way interaction (model E), respectively.

2.5.2 Design factors

Inspired by previous simulation studies by Viechtbauer (2007b) and Higgins and Thompson (2004), the influence on recovery performance of five design factors was examined: (a) the number of studies (K); (b) the average within-study sample size ($\bar{n}$); (c) the residual heterogeneity (σ_τ^2); (d) the number of study-level covariates (M), and (e) the magnitude of the interaction effect (δ_I). For each true model with each combination of the design factors, 1000 datasets were generated and analyzed using each of the options for each type of tree, implying 9×2 versions of meta-CART.

In a pilot simulation study, it was found that all versions of meta-CART applied to data sets with $K = 20$ studies result in poor power rates (≤ 0.30). Therefore, $K = 20$ was not included in our final simulation. Instead, three values of K were chosen: 40, 80, 120.

By adjusting the value of the within-study sample size n_k, the amount of sampling error $\sigma_{\epsilon_k}^2$ can be manipulated. We used the same method as in Viechtbauer (2007b) to generate n_k, by which the values of n_k were sampled from a normal distribution with mean $\bar{n}$ and standard deviation $\bar{n}/3$. Three levels of the average within-study sample size $\bar{n}$ were chosen as 40, 80, 160. The resulting n_k ranged roughly between 15 and 420, which are values encountered in practice.

When searching for the combinations of covariates (e.g., BCTs) that result in the most effective interventions, the covariates are usually coded as binary variables, that is, 0 for "not included", and 1 for "included". Therefore, in the simulation study we focused on the detection of interaction effects between binary moderators. To assess how many covariates (i.e., potential moderators) meta-CART can deal with to successfully identify the true moderators and the interaction effect(s) between them, three values of the number of covariates M were chosen: 5, 10, 20.

Unfortunately, the values of residual heterogeneity are barely reported in the literature on meta-analysis. In a very few papers that reported values of residual heterogeneity, σ_τ ranges are between 0 and 0.05 (Dusseldorp et al., 2014; Viechtbauer, 2007a). Thus, the values of σ_τ^2 were chosen as 0, 0.025, 0.05.

The magnitude of the interaction effect was measured by the average effect size of the studies in the effective subgroups δ_I. The pilot simulation study showed that all versions of meta-CART failed to achieve enough power (≤ 0.70) to detect a small interaction effect with $\delta_I = 0.2$. Therefore, two values of δ_I were chosen as 0.5 and 0.8, corresponding to a medium and a large effect size, respectively (Cohen, 1988).

2.5.3 Monte Carlo Simulation

Artificial data sets were generated from each true model with each combination of the design factors. As mentioned before, in each cell of the design, 1000 meta-analytic data

sets were generated. Each meta-analytic data set consists of two subsets of the same size: a training data set for fitting the model, and a test data set for estimating the prediction error.

Within each data set, binary study characteristics were generated independently from a Bernoulli distribution with a probability of 0.50. For a single study, the true effect size δ_k was sampled from a normal distribution with mean Δ and variance σ_τ^2, where Δ is the average population effect size. The value of Δ depends on the moderators and the corresponding true model. Then the observed effect size g_k was sampled from a non-central t-distribution (see Appendix A).

The 9×2 versions of meta-CART were applied to each generated data set. The interaction effects were investigated in the first step, and the significance was tested by the between subgroups Q-statistic with $\alpha = 0.05$ in the second step on the same data set.

2.5.4 The evaluation criteria for success

Three criteria are employed to judge whether meta-CART successfully retrieved the true model underlying the data:

Criterion 1. Meta-CART correctly detects the presence of moderator effect(s) in the data sets generated from model B, C, D or E (power).

Criterion 2. Meta-CART obtains a pruned tree with exactly the same number of terminal nodes as the true structure underlying the data (recovery of tree complexity).

Criterion 3. Meta-CART successfully selects the study characteristics used in the true model (recovery of moderator(s)).

Each of the three criteria was evaluated and coded with 0 for "not satisfied" and 1 for "satisfied" for each data set. Subsequently, for each cell of the design, the proportion of successful (i.e., "satisfied") solutions was computed per criterion.

2.6 Results

2.6.1 The best options on average

The first goal of the simulation study was to find the best combination of options of each type of meta-CART that has the best overall performance in most conditions (i.e., across different design factors). The performance of meta-CART was evaluated in terms of the Type I error rates and the three evaluation criteria. Table 2.2 shows the estimated Type I error rates averaged over all design factors for the 9×2 versions of meta-CART. The average Type I error rates of meta-regression trees range from .014 to .118, and the

Table 2.2: Type I error rates of meta-CART, averaged over design factors.

		Meta-regression tree						Meta-classification tree		
model	c	$\mathcal{W}_0$		$\mathcal{W}_1$		$\mathcal{W}_2$		$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$
A	0.0	**.114**	**(.029)**	**.118**	**(.028)**	**.117**	**(.029)**	**.485 (.179)**	**.447 (.173)**	**.477 (.189)**
	0.5	.041	(.021)	.044	(.022)	.042	(.022)	**.447 (.156)**	**.398 (.141)**	**.432 (.159)**
	1.0	.014	(.013)	.015	(.014)	.014	(.013)	**.391 (.130)**	**.347 (.112)**	**.380 (.132)**

Table 2.3: The power rates of meta-CART, averaged over models B, C, D, E and design factors.

	Meta-regression tree			Meta-classification tree		
c	$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$	$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$
0.0	.916 (.167)	.912 (.171)	.916 (.167)	.948 (.093)	.929 (.114)	.941 (.101)
0.5	.881(.213)	.877 (.217)	.881 (.213)	.941 (.101)	.917 (.130)	.929 (.114)
1.0	.846 (.251)	.843 (.254)	.847 (.251)	.928 (.118)	.895 (.150)	.914 (.131)

Table 2.4: The recovery rates of tree complexity, averaged over models B, C, D, E and design factors.

	Meta-regression tree			Meta-classification tree		
c	$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$	$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$
0.0	.688 (.242)	.681 (.243)	.685 (.242)	.522 (.375)	.514 (.368)	.519 (.370)
0.5	.752(.304)	.747 (.305)	.754 (.302)	.521 (.382)	.527 (.379)	.521 (.380)
1.0	.750 (.339)	.747 (.340)	.753 (.338)	.509 (.394)	.506 (.389)	.511 (.391)

Table 2.5: The recovery rates of moderators, averaged over models B, C, D, E and design factors.

	Meta-regression tree			Meta-classification tree		
c	$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$	$\mathcal{W}_0$	$\mathcal{W}_1$	$\mathcal{W}_2$
0.0	.813 (.307)	.810 (.312)	.814 (.309)	.612 (.369)	.618 (.367)	.624 (.365)
0.5	.784(.334)	.781 (.337)	.785 (.359)	.579 (.381)	.577 (.381)	.581 (.381)
1.0	.757 (.352)	.755 (.353)	.759 (.351)	.537 (.396)	.538 (.393)	.541 (.394)

standard deviations range from .013 to .029. The averaged Type I error rates of meta-classification trees range from .347 to .485 and the standard deviations range from .112 to .189. In general, the Type I error rates mainly depend on the type of tree (meta-regression trees outperform meta-classification trees) and the pruning rule, but not on the type of weights.

Tables 2.3, 2.4 and 2.5 show the power rates, the recovery rates of tree complexity, and the recovery rates of moderators averaged over all design factors for the 9×2 versions of meta-CART, respectively. The three criteria were averaged over the models B, C, D and E since the patterns are similar in these four models. Again, it was found that the variation of the three criteria mainly depends on the type of the tree and the pruning rule, but not on the type of weights. In general, meta-classification trees result in higher power rates than meta-regression trees. But meta-regression trees outperformed meta-classification trees in terms of the two recovery rates. For meta-regression trees, applying weights $\mathcal{W}_2$ result in slightly better performance than applying weights $\mathcal{W}_0$ or $\mathcal{W}_1$ in terms of all three criteria. We chose the best combinations of options based on the Type I error rates and the three criteria. An average Type I error below .05 was chosen to be acceptable in order to control for the risk of finding spurious interaction effects. As a result, there is no good combination of options for meta-classification trees. For meta-regression trees, applying weights $\mathcal{W}_2$ and a pruning rule with $c = 0.5$ was chosen as the best combination. With control of acceptable Type I error rates, this combination of options has the highest power rates, recovery rates of tree complexity and recovery rates of moderators

2.6.2 The influence of the design factors

The second goal of the simulation study is to evaluate the influence of the design factors on the Type I error and the three criteria. We focus on the results of meta-regression trees using only the best combination of options (as defined above). The Type I error rates and the three criteria were computed separately for each model and each combination of the design factors. The resulting proportions were subjected to an analysis of variance (ANOVA) with the five design factors and their interactions as independent variables and the five-way interaction being used as an error term.

For model A, the ANOVA results reveal that only the number of studies (K) has a strong influence (partial $\eta^2 > 0.80$) on the Type I error rates. The Type I error rates decrease with increasing K (.069, .035, .021 when $K = 40$, 80, 120, respectively).

For models B, C, D, and E, the ANOVA results reveal that all the design factors and most of their interactions have strong influence (partial $\eta^2 > 0.80$) on the three criteria. Despite some noise, in general, the three criteria are positively related to the number of studies (K), the average within-study sample size ($\bar{n}$), and the magnitude of interaction effects (δ_I). On the other hand, the three criteria are negatively influenced by the number

of potential moderators (M) and the residual heterogeneity (σ_τ^2). The plots representing the three criteria have similar patterns. Because the recovery rates of tree complexity is the most strict criterion among the three, these plots are represented in Figures 2.3, 2.4 and 2.5 for each cell of the design. The plots representing the power rates and the recovery rates of moderators can be found in Supporting Materials. Note that the Figures are ordered from $K = 120$ to $K = 40$ (left to right).

When there is only one main effect in the true model (model B), the recovery rates are satisfactory ($\geq .80$; not shown) in all cases, except one: it is .75 in the case of a medium-sized main effect, a small number of studies,a small within-study sample size, a large number of moderators and large residual heterogeneity ($K = 40, \bar{n} = 40, \delta_I = 0.8, M = 20, \sigma_\tau^2 = 0.05$).

Figure 2.3 represents the recovery rates of tree complexity of meta-regression trees when there is one two-way interaction in the true model (model C). When $K = 120$, meta-regression trees can always achieve a satisfactory recovery rate ($\geq .80$). When $K = 80$, the recovery rates are satisfactory in most cases, with one exception: the recovery rates to detect a medium-sized interaction effect ($\delta_I = 0.5$) are between .60 and .80 in cases of large residual heterogeneity and a large number of study characteristics ($\sigma_\tau^2 = 0.05$, $M = 20$). When $K = 40$, the average within-study sample size needs to be large enough ($\bar{n} \geq 160$) to achieve a satisfactory recovery rate in cases of no residual heterogeneity (Figure 2.3c). In cases of residual heterogeneity (Figures 2.3f, 2.3i), meta-regression trees can achieve a satisfactory recovery rate for a medium-sized interaction effect only if the residual heterogeneity is relatively small ($\sigma_\tau^2 = 0.025$), the within-study sample size is large ($\bar{n} = 160$), and the number of moderators is small ($M = 5$). If the average within-study sample size is large enough and the number of study characteristics is relatively small ($\bar{n} \geq 80$, $M \leq 10$), the recovery rates are satisfactory for a large-sized interaction effect ($\delta_I = 0.8$).

Figure 2.4 represents the recovery rates of tree complexity of meta-regression trees when there are two two-way interactions in the true model (model D). When $K = 120$, meta-regression trees achieve satisfactory recovery rates in most cases, with four exceptions: the recovery rates for a medium-sized interaction effect are between .5 and .8 in the case of residual heterogeneity ($\sigma_\tau^2 \geq 0.025$) and $M \geq 10$ (Figure 2.4g). When $K = 80$, the picture is more complex. The recovery rates for detection of large-sized interaction effects are satisfactory. If $\bar{n} \geq 160$, the recovery rate is satisfactory to detect a medium-sized interaction effect in case of small residual heterogeneity (Figure 2.4e). In case of large residual heterogeneity (Figure 2.4h) meta-regression trees fail to achieve satisfactory recovery rates to detect medium-sized interaction effects. When $K = 40$, meta-regression trees fail to achieve satisfactory recovery rates.

Figure 2.5 represents the recovery rates of tree complexity of meta-regression trees when there is one three-way interaction in the true model (model E). When $K = 120$, the

recovery rates for a large-sized interaction effect are satisfactory. To achieve a satisfactory recovery rate for detection of medium-sized interaction effect, the average within-study sample size needs to be large ($\bar{n} \geq 160$). In addition, the number of study characteristics also needs to be small ($M \leq 5$) in the case of large residual heterogeneity ($\sigma_\tau^2 = 0.05$). When $K = 80$, $\bar{n}$ needs to be large enough (≥ 160) to achieve satisfactory recovery rates in the case of no residual heterogeneity (Figure 2.5b). In cases of residual heterogeneity (Figures 2.5e, 2.5h), meta-regression trees fail to achieve satisfactory recovery rates for detection of medium-sized interaction effects. For detection of large-sized interaction effects, the within-study sample size needs to be large enough ($\bar{n} \geq 80$) for meta-regression trees to achieve satisfactory recovery rates. In addition, the number of moderators needs to be small ($M \leq 10$) in case of large residual heterogeneity. When $K = 40$, meta-regression trees fail to achieve satisfactory recovery rates.

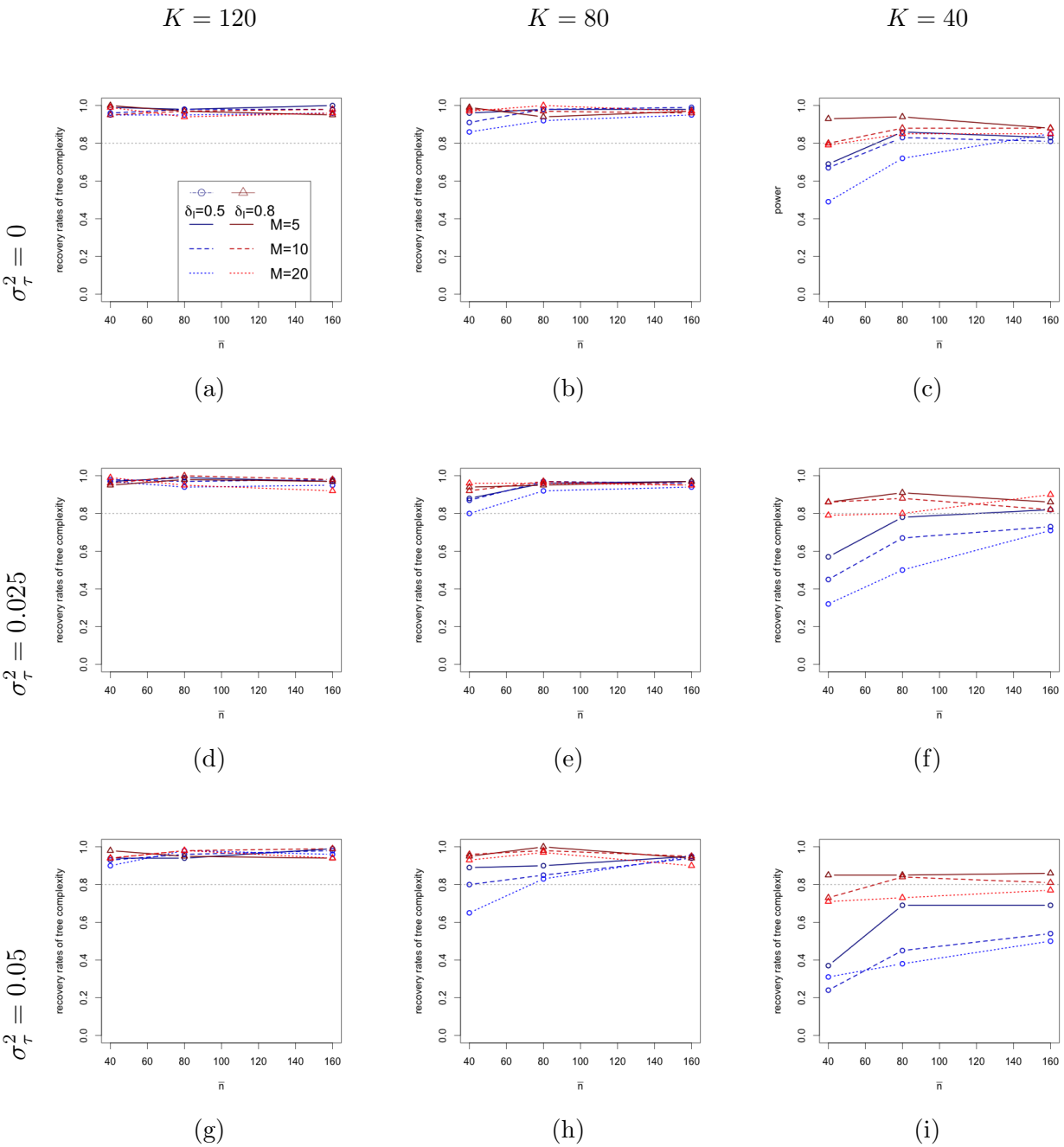


Figure 2.3: Recovery rates of tree complexity(y-axis) of meta-regression trees for model C. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ^2_τ). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

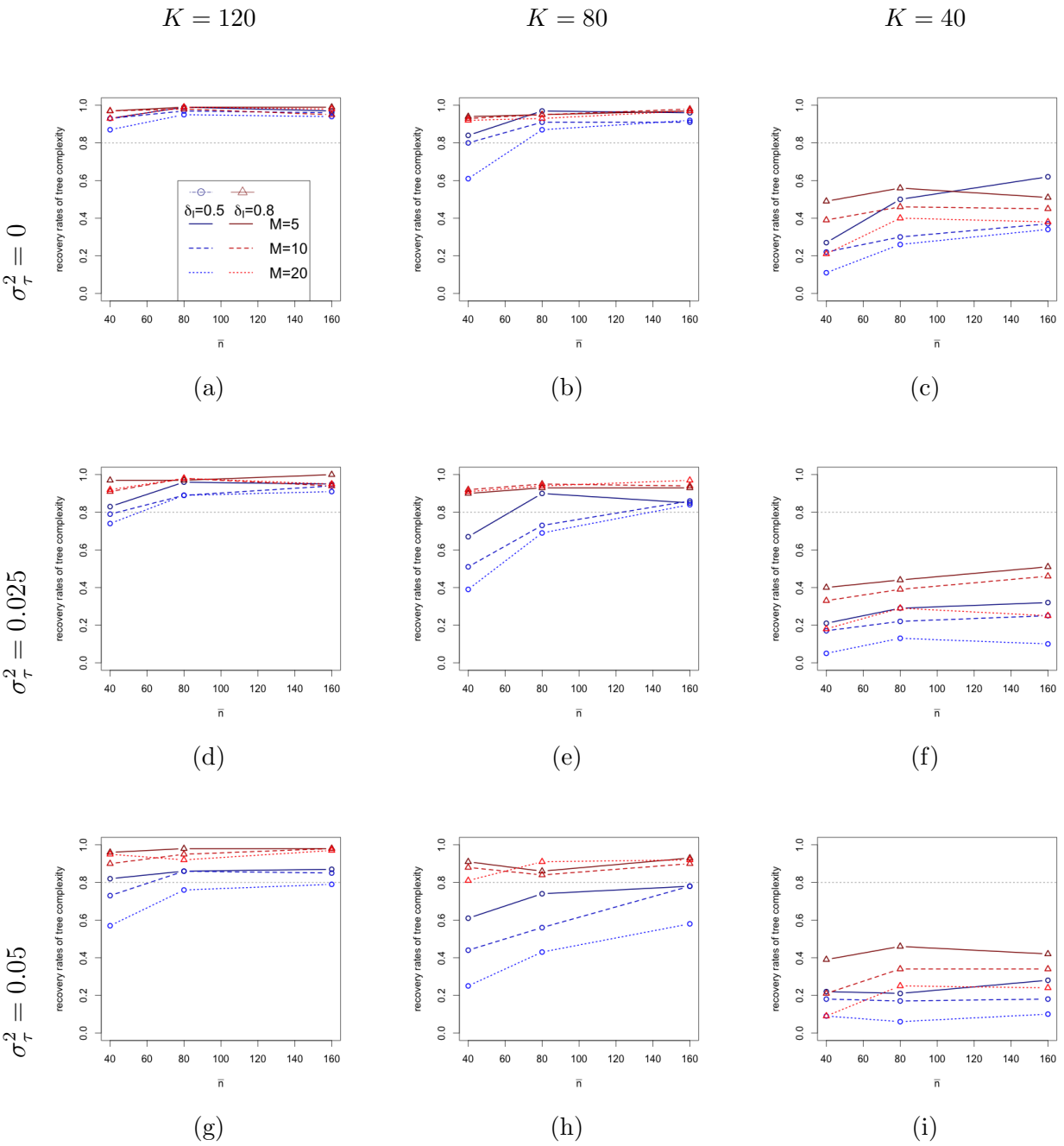


Figure 2.4: Recovery rates of tree complexity (y-axis) of meta-regression trees for model D. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ^2_τ). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

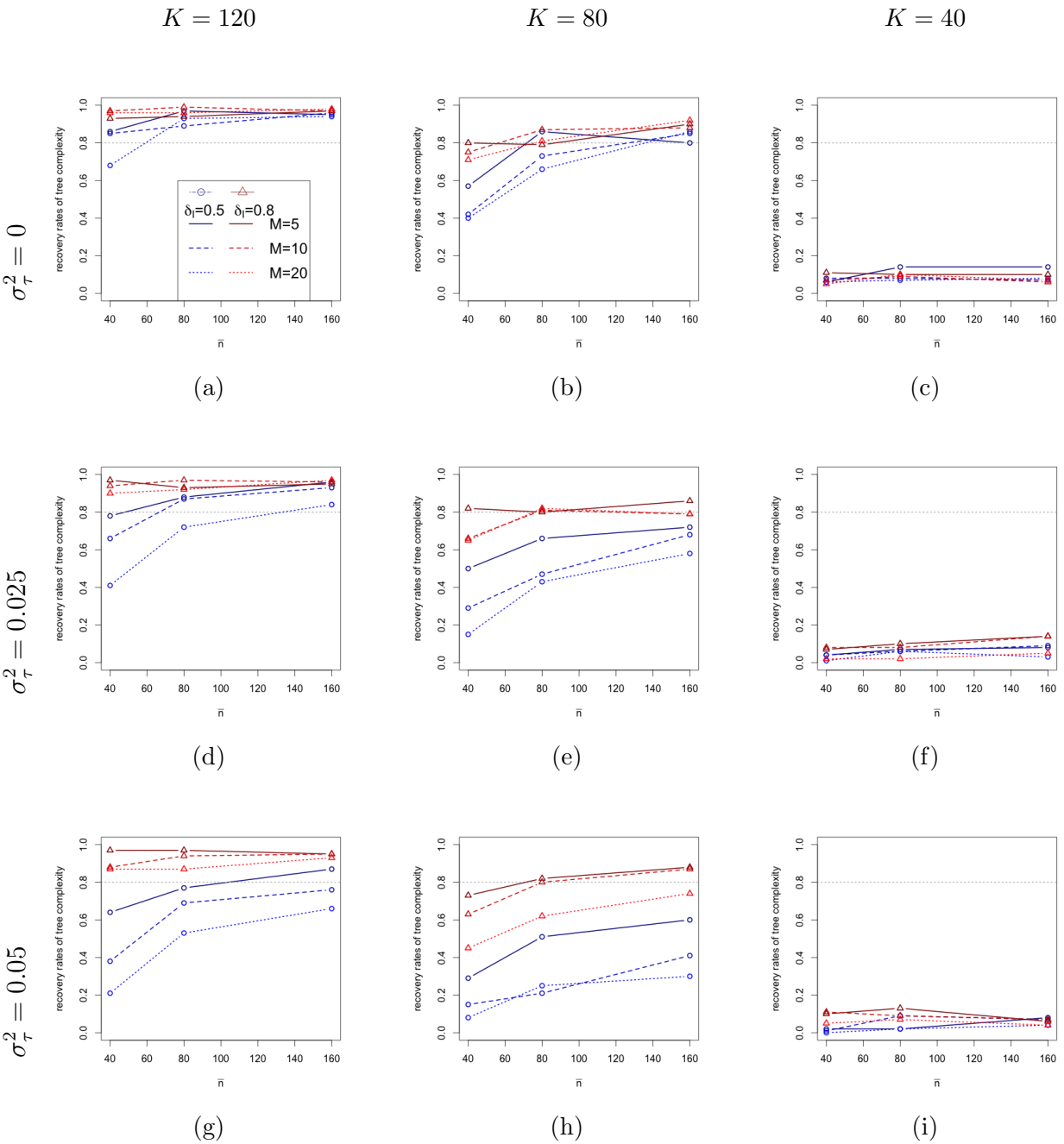


Figure 2.5: Recovery rates of tree complexity (y-axis) of meta-regression trees for model E. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ^2_τ). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

2.7 Discussion

2.7.1 General discussion

The present study proposed extensions for the meta-CART approach by Dusseldorp et al. (2014) and investigated the performance of the previous and the extended options for meta-CART via an extensive simulation study. The previous version of meta-CART considered meta-classification trees only, whereas in this study meta-regression trees were also used. Three options were defined for the weights (i.e., applying no weights, fixed effects weights, or random effects weights), and three options for the pruning rules (i.e., $c = 0$, 0.5 or 1). The first question of the simulation study was directed at finding the best combinations of options for meta-classification trees and meta-regression trees separately. The simulation results show that meta-regression trees have the best performance on average when applying random effects weights and using a pruning rule with $c = 0.5$. Furthermore, no best combination of options could be found on average for meta-classification trees, because all options failed to control the Type I error rate. The second question focused on exploring the influence of the design factors on the Type I error rates, the power rates, the recovery rates of tree complexity, and the recovery rates of moderators. The results revealed that the performance of meta-regression depends on all the design factors (i.e., the number of studies, the within-study sample size, the number of moderators, the magnititude of interaction effect size, and the residual heterogeneity) and the complexity underlying the data. These results were used to formulate guidelines for application (see Section 2.7.3).

Meta-CART was proposed to overcome the difficulties of standard meta-regression to deal with multiple study characteristics and higher-order interactions in exploratory meta-analytic studies. However, if there are a priori hypotheses about possible interactions between moderators (i.e., in confirmatory studies), the advantage of standard meta-regression is that the interaction(s) can be specified beforehand and tested. Thus, meta-regression is the recommended tool for confirmatory meta-analytic studies. On the other hand, meta-CART is recommended for exploratory studies, especially for studies with multiple study characteristics and for studies interested in higher-order interactions.

Another difference between meta-CART and meta-regression is the interpretation of the interaction effects. The interaction effects in meta-CART are presented in a parsimonious tree structure. In meta-regression, interaction effects are represented by strictly additive functions. For example, model D in Figure 2.2 with a medium-sized interaction effect can be expressed as

$$\delta_k = 0.5 \cdot x_{3k} + 0.5 \cdot x_{1k}x_{2k} - 0.5 \cdot x_{1k}x_{3k}. \tag{2.5}$$

It is worth noting that, without looking at the interactions, researchers might draw the

false conclusion that interventions with the characteristic x_3 have a positive treatment effect in general. Compared to a linear regression model, the tree model directly shows that the intervention is effective only in those interventions with x_1 and x_2 both present, and those with x_3 present but x_1 absent. Although the tree structure provides a straightforward visual representation with easy interpretability, a downside of meta-CART may be that main effects and interaction effects between multiple moderators are hardly distinguished. In contrast, regression models are well suited for representing strictly additive functions but may not be able to represent complex interaction patterns and nonlinear effects (Little, 2013).

2.7.2 Strengths, shortcomings, and remaining issues

One strength of our study is that the design factors of the simulation covered most values that have been encountered in practice. Our results show that the power rates, the recovery rates of tree complexity and the recovery rates of moderators are well discriminated by the different values of design factors, that is, the conditions resulting in high performance in the three criteria and those resulting in low performance are both encountered.

There are still some shortcomings in our study, some concerning the simulation design and others concerning the meta-CART algorithm. First, the potential moderators in the simulation study only contain binary variables, and were independently generated. For binary moderators, there is only one possible split point for each moderator. As a result, meta-regression trees were found to be stable in the simulation study. However, the stability of meta-regression trees might be an issue when dealing with nominal or continuous moderator variables, since there are more possible split points. Second, the true models that we designed to generate the data sets contained only one or two interaction effects between moderators. However, interactions can be much more complex in real-world data (e.g., Dusseldorp et al., 2014). Furthermore, the true intervention effect sizes in the ineffective subgroups were designed to be all 0s, and the true effect sizes for the effective subgroups only contained two values: 0.5 or 0.8. This might be too simplistic. We intend to investigate the performance of meta-CART on nominal or continuous predictors and more complex scenarios in future work.

There are also some shortcomings concerning the meta-CART algorithm. First, in our pilot simulation study, it was found that all implementations of meta-CART had limited detection rates (≤ 0.30) when applied to data sets with $K = 20$ studies. This means that meta-CART is not recommended for meta-analytic data with $K \leq 20$ studies. Second, the procedure follows a step-wise approach. As a result, it lacks efficiency, and it uses a local optimization procedure, which is a general shortcoming of recursive partitioning methods. The two-step procedure also raises the issue of statistical inference. It would be interesting in future work to integrate the two steps, and to investigate whether a

global optimization procedure is possible. One possible solution is to maximize a test statistic, for example, the chi-square statistic, over all possible combinations of covariates (see Boulesteix, 2006). Another possible solution is to implement a statistical test within the recursive partitioning to determine the best size of the tree (Hothorn, Hornik, & Zeileis, 2006). Third, the improvement in the performance of meta-CART by applying random-effects weights was small. One possible reason is that the current partitioning criterion is based on the CART algorithm, and it does not take into account the residual heterogeneity (i.e., σ_τ^2). A possible improvement might be applying random-effects weights together with a partitioning criterion that maximizes the between subgroups Q-statistic or minimizes the residual heterogeneity in meta-CART.

The simulation results revealed that all options for meta-classification trees result in high average Type I error rates (ranging from 0.347 to 0.485). A possible explanation is that the dichotomization of the *response* variable (i.e., study effect sizes) in classification trees results in spurious results and inflates the Type I error rates. In Maxwell and Delaney (1993), it has been shown that dichotomizing multiple *predictor* variables may dramatically increase the probability of Type I errors in some situations. This study shows that dichotomizing the response variable may also increase this probability. Another reason of the Type I error inflation is that the subgroup meta-analysis in the second step does not hold the nominal alpha level. Meta-CART detects and then tests interaction effects between moderators by using the same data set. Such subgroup meta-analyses are post-hoc tests and raise problems of statistical inference. In the two-step procedure, the Type I error rates are controlled by two things: the pruning parameter and the nominal alpha level of the subgroup Q statistic. To control the Type I error, we fixed the nominal alpha level 0.05 since it is the most commonly used in practice, and we chose the best pruning rule that results in acceptable Type I error rates. For meta-regression trees, this approach finds $c = 0.50$ as the best pruning parameter. For meta-classification trees, however, no pruning parameter can prevent the inflation of Type I error.

2.7.3 Guidelines for application of meta-CART

According to the simulation results, we recommend to use meta-regression trees instead of meta-classification trees. Furthermore, applying random-effects weights and a pruning rule with $c = 0.50$ was chosen as the best combination of options for meta-regression trees. There is, however, a shortcoming for applying random-effects weights in a practical application: the computation of the random-effects weights requires the value of the residual heterogeneity $\sigma_{\tau_r}^2$, which is not known a priori. There are two possible solutions for this problem. One suggested solution is to apply a meta-regression tree without weights and a pruning rule with $c = 0.50$ instead, since the simulation study shows that there is only a slight difference in the performance between the two combinations of options.

An alternative solution is to first estimate the $\sigma^2_{\tau_r}$ by employing a meta-regression tree without weights, and then to use the estimated residual heterogeneity $\hat{\sigma}^2_{\tau_r}$ to compute the random-effects weights.

Based on the simulation study results, recommendations can be made about the number of studies included in a meta-analysis using meta-CART to achieve a satisfactory performance. This number depends on the complexity of the data and the number of study characteristics. In general, 40 is the minimum number of studies required for meta-CART to perform well in detecting simple interaction effects, that is, only one two-way interaction. To detect more complex interaction effects, such as more than one two-way interactions, or higher-order interactions, at least 80 studies are needed for meta-CART to achieve a power higher than 0.80. Data sets with 120 or more studies would be ideal since meta-CART has good performance in most cases, even in cases with residual heterogeneity or complex interaction effects. In the case of large residual heterogeneity (≥ 0.025) and complex interaction effects, meta-CART requires more studies and larger within-study sample size ($K \geq 120$, $\bar{n} \geq 160$) for data sets with more than five study characteristics. However, the performance of meta-CART is not much influenced by the number of study characteristics in the case of small residual heterogeneity or a two-way interaction effect.

2.8 Supporting Materials

2.8.1 Plots displaying the power rates

Figures 2.6, 2.7, and 2.8 represent the power rates of meta-regression trees for models C, D, and E, respectively. In general, the power rates are positively related to the number of studies (K), the average within-study sample size ($\bar{n}$), and the magnitude of interaction effects (δ_I), except for some noise (e.g., 2.6i). On the other hand, the power rates are negatively influenced by the number of potential moderators (M) and the residual heterogeneity (σ^2_τ).

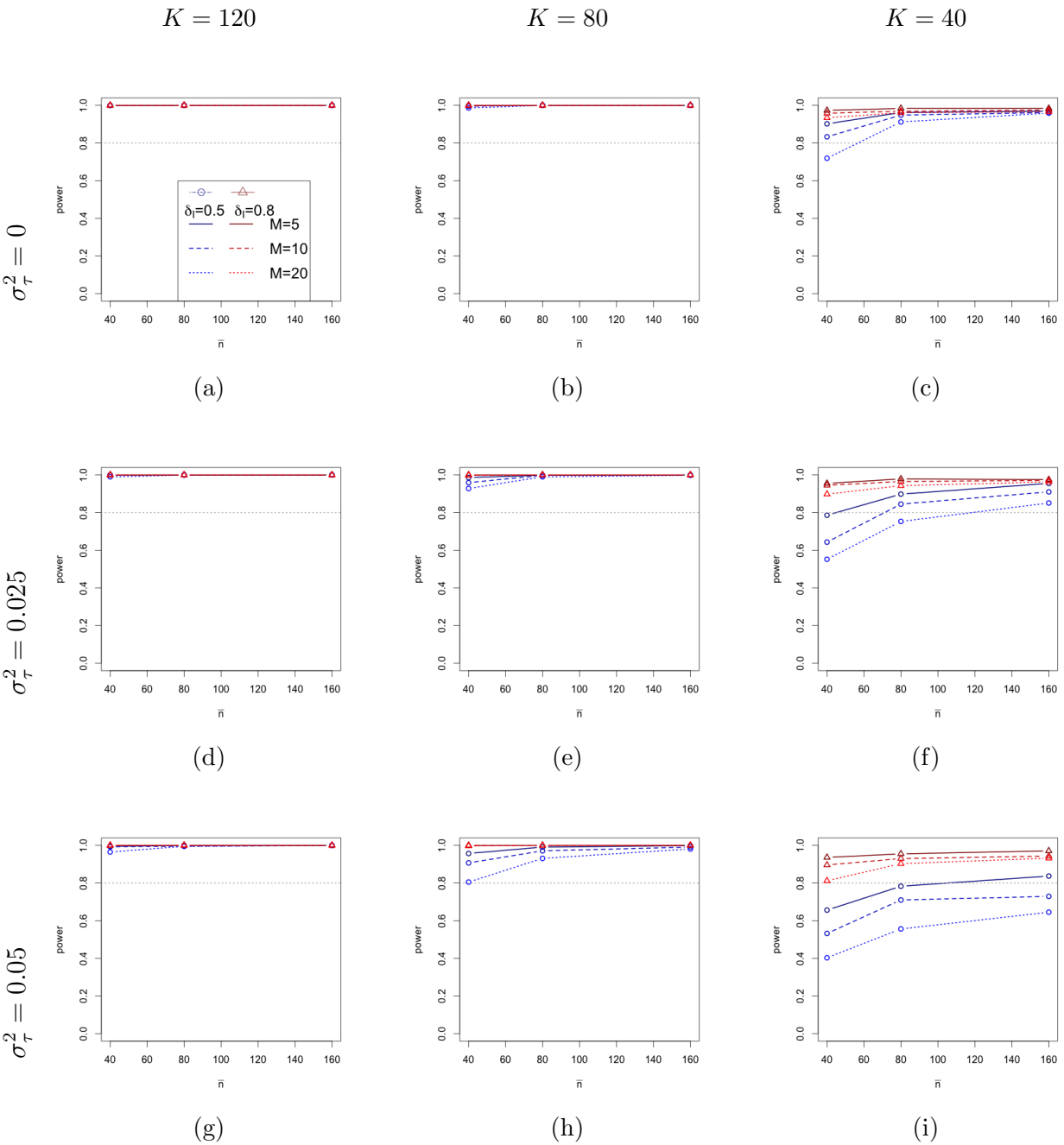


Figure 2.6: Power rates (y-axis) of meta-regression trees for model C. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ^2_τ). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

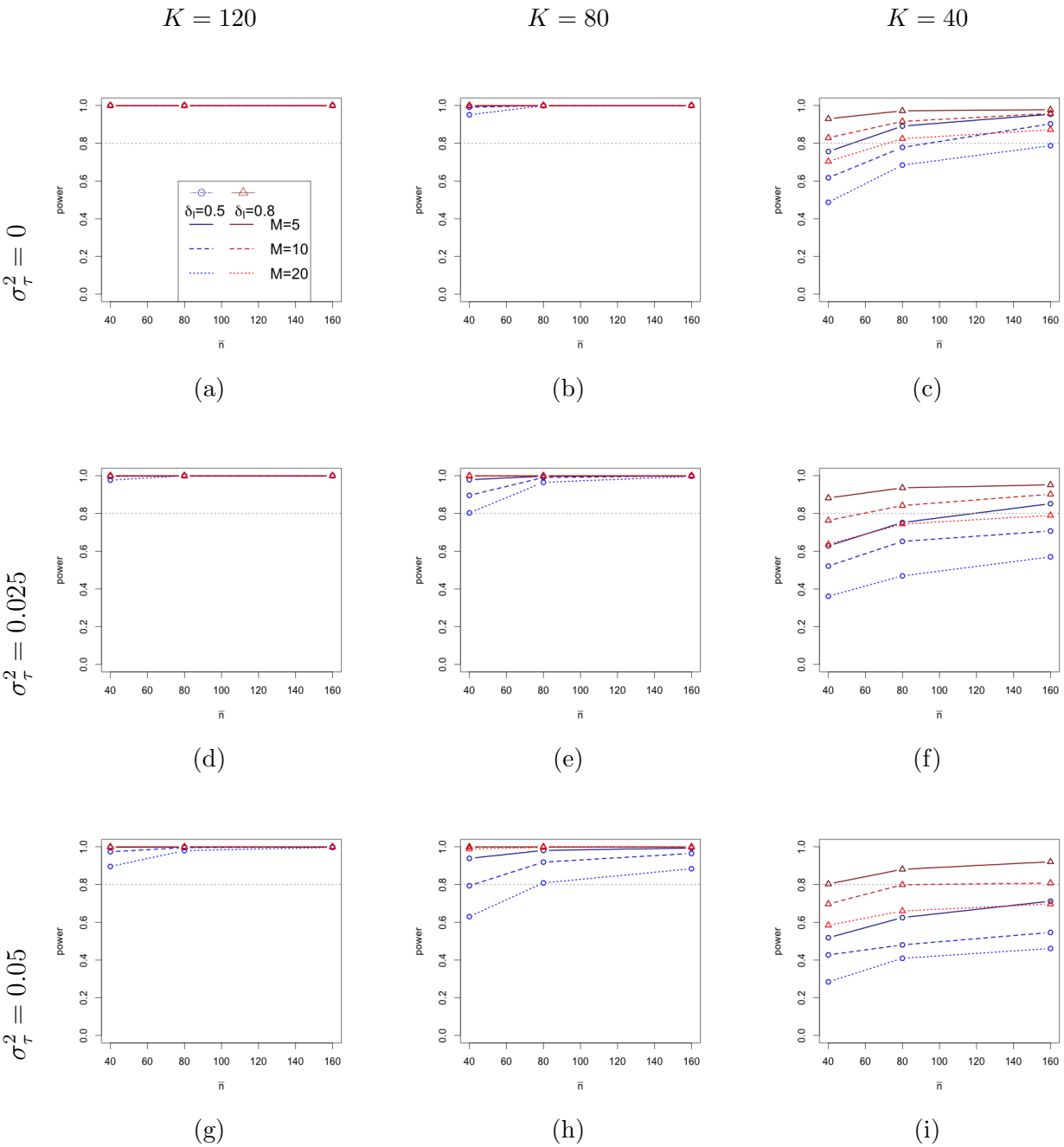


Figure 2.7: Power rates (y-axis) of meta-regression trees for model D. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ_τ^2). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

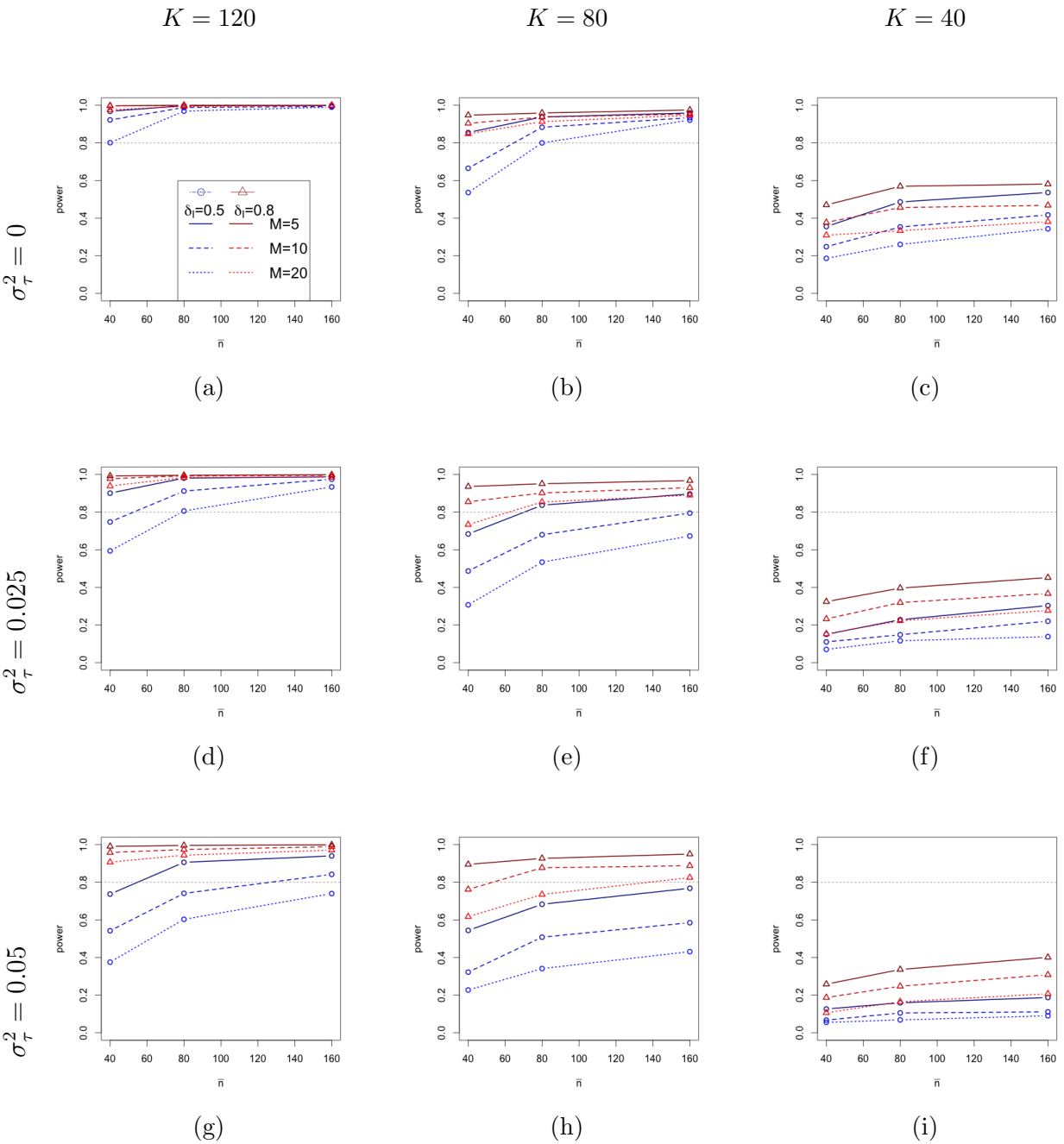


Figure 2.8: Power rates (y-axis) of meta-regression trees for model E. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ_τ^2). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

2.9 Plots displaying recovery rates of moderators

Figures 2.9, 2.10, and 2.11 represent the recovery rates of moderators of meta-regression trees for model C, D, and E, respectively. The plot is not shown for model B because the recovery rates of moderators are higher than 0.80 in all cases. Similar to the recovery rates of tree complexity, there is noise when the recovery rates are relatively low, or the interaction terms in the true model are complex (i.e., model E). In general, the recovery rates of moderators are positively related to the number of studies (K), the average within-study sample size ($\bar{n}$), and the magnitude of interaction effects (δ_I). On the other hand, the recovery rates of moderators are negatively influenced by the number of potential moderators (M) and the residual heterogeneity (σ^2_τ).

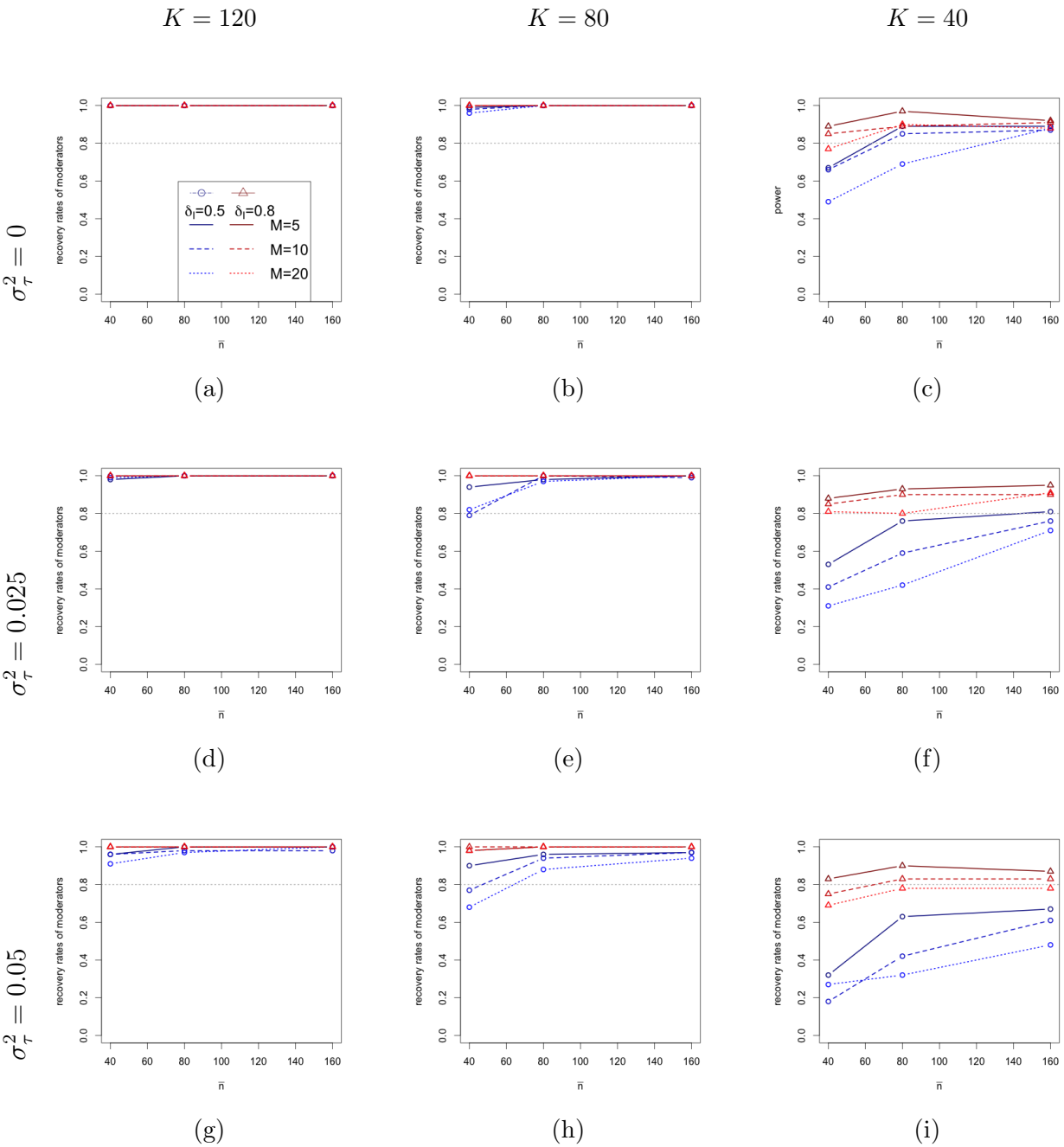


Figure 2.9: Recovery rates of moderators (y-axis) of meta-regression trees for model C. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ_τ^2). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

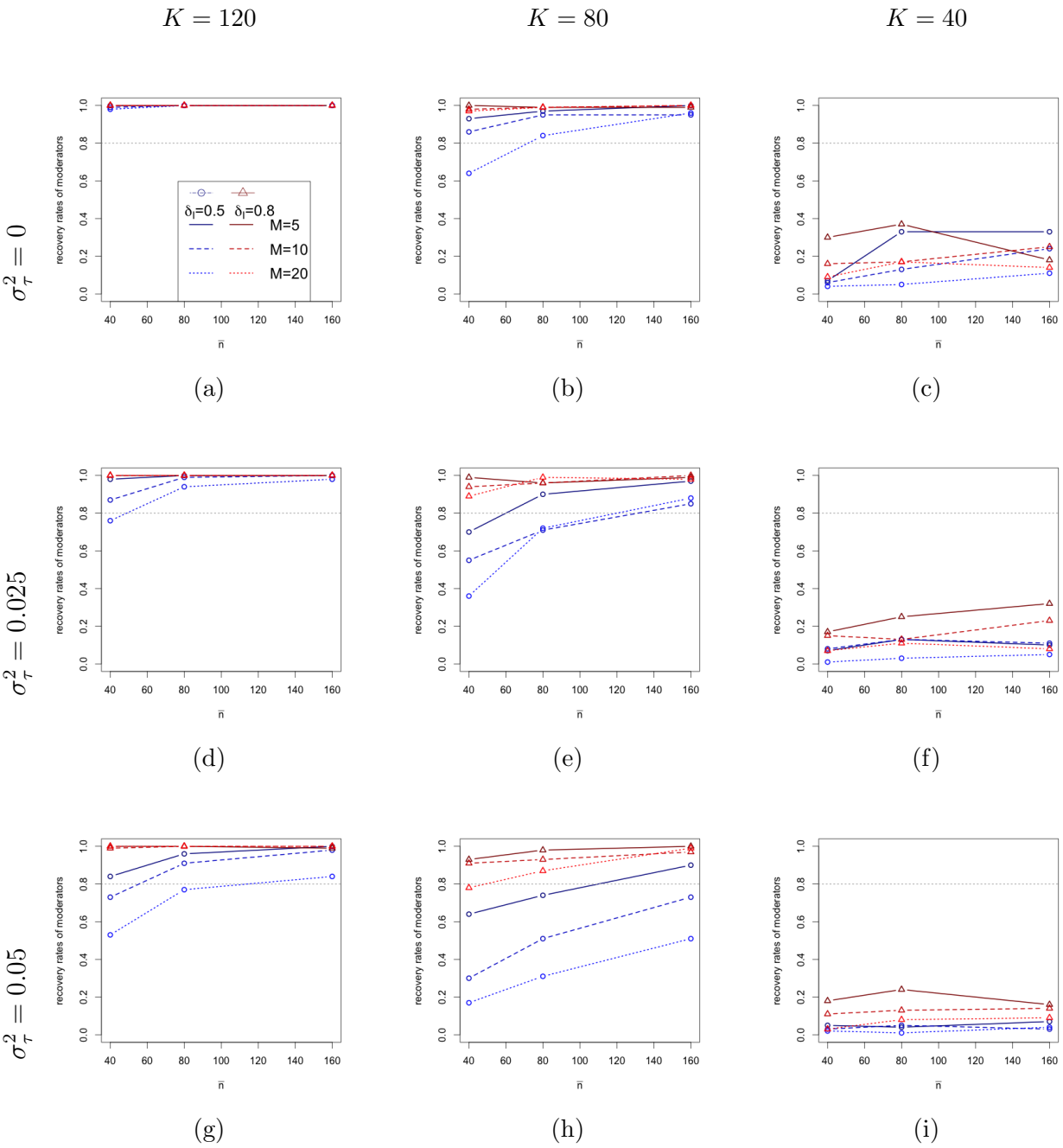


Figure 2.10: Recovery rates of moderators (y-axis) of meta-regression trees for model D. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ_τ^2). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

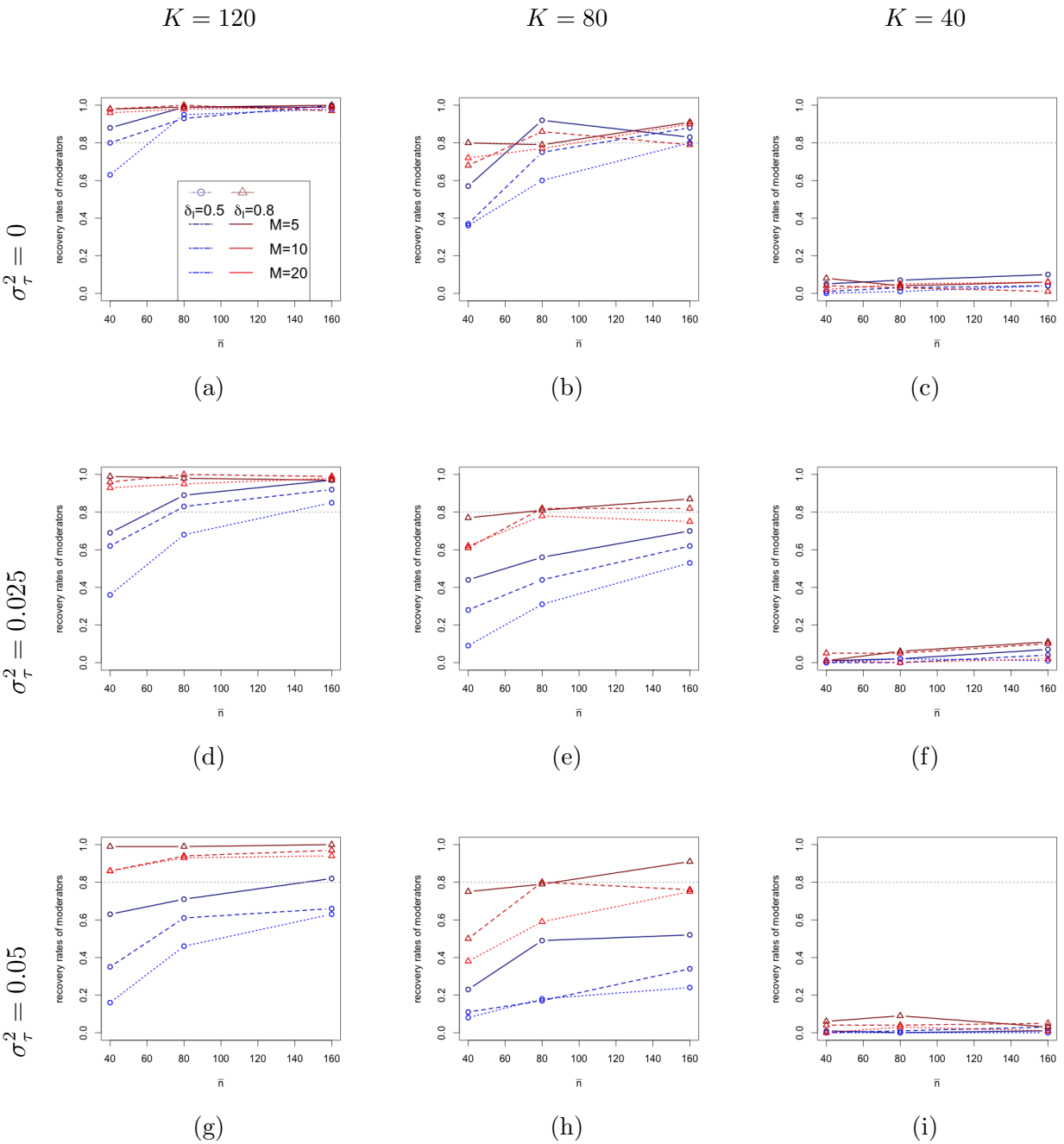


Figure 2.11: Recovery rates of moderators (y-axis) of meta-regression trees for model E. Separate plots are shown for the number of studies (K), the average within-study sample size (n, x-axis) and the residual heterogeneity (σ_τ^2). Separate lines are shown for the combination of the number of moderators (M) and the interaction effect size (δ_I). The legends of all plots are shown in plot (a).

Chapter 3

A flexible approach to identify interaction effects between moderators in meta-analysis

abstract

In meta-analytic studies, there are often multiple moderators available (eg, study characteristics). In such cases, traditional meta-analysis methods often lack sufficient power to investigate interaction effects between moderators, especially high-order interactions. To overcome this problem, meta-CART was proposed: an approach that applies classification and regression trees (CART) to identify interactions, and then subgroup meta-analysis to test the significance of moderator effects. The aim of this study is to improve meta-CART upon two aspects: 1) to integrate the two steps of the approach into one and 2) to consistently take into account the fixed-effect or random-effects assumption in both the the interaction identification and testing process. For fixed effect meta-CART, weights are applied, and subgroup analysis is adapted. For random effects meta-CART, a new algorithm has been developed. The performance of the improved meta-CART was investigated via an extensive simulation study on different types of moderator variables (i.e., dichotomous, nominal, ordinal, and continuous variables). The simulation results revealed that the new method can achieve satisfactory performance (power greater than 0.80 and Type I error less than 0.05) if appropriate pruning rule is applied and the number of studies is large enough. The required minimum number of studies ranges from 40 to 120 depending on the complexity and strength of the interaction effects, the within-study sample size, the type of moderators, and the residual heterogeneity.

3.1 Introduction

The primary aims of meta-analysis are to synthesize the estimates of an effect or outcome of interest from multiple studies (i.e., effect size), and to assess the consistency of evidence among different studies (i.e., heterogeneity test). When study features (i.e., moderators) are available, meta-analysis can be used to assess the influence of the study features on the study outcomes. In recent years, there is a growing need to integrate research findings due to the increasing number of publications. As research questions and data structures are becoming more complex, there are often multiple moderators involved in meta-analytic data (e.g., Michie, Abraham, et al., 2009). In such cases, conventional univariate meta-analytic techniques (e.g., Hedges & Olkin, 1985; Schmidt & Hunter, 2014) may not be appropriate. Multivariate meta-analytic techniques, for example, meta-regression, are required to assess the influence of multiple moderators on the effect size.

When multiple moderators are available, the effects of moderators may be non-additive, and the moderators may attenuate or amplify each other's effect. In such situations, interaction effects between moderators occur. Knowledge about interaction effects may provide valuable information. For example, when treatment alternatives consist of several components, the researchers might be interested in questions such as "Which combination of components is most effective?" (Welton et al., 2009). Knowledge of effective combinations can be helpful to evaluate existing treatments (e.g., by examining whether an effective combination of treatment components is used) and to design new potentially effective treatments (e.g., by choosing the effective combinations of treatment components).

Despite the need to investigate multiple moderator variables and the interaction effects between them, most meta-analytic studies apply univariate moderator analyses only (e.g., Huisman et al., 2009; Yang & Raine, 2009). And even in studies employing multivariate meta-analytic techniques, interaction effects were seldom investigated. Possible reasons are the lack of appropriate methods and corresponding software for identifying interaction effects in meta-analyses. To solve this problem, a new strategy, called meta-CART, which integrates classification and regression trees (CART; Breiman et al., 1984) into meta-analysis, was proposed (Dusseldorp, van Genugten, van Buuren, Verheijden, & van Empelen, 2014; Li, Dusseldorp, & Meulman, 2017). This method can deal with many predictors and represents interactions in a parsimonious tree structure. The results of meta-CART were promising from a substantial point of view (Dusseldorp et al., 2014), that is, the method could produce interpretable and meaningful results for real-world data. Also, meta-CART has the potential to be an alternative statistical method for meta-regression to understand the combined effects of moderators (Michie et al., 2015; O'Brien et al., 2015). The results of a previous simulation study (Li, Dusseldorp, & Meulman, 2017) showed that, regression trees in meta-CART have better performance than classification trees. Meta-CART achieved satisfactory power and recovery rates

(i.e., ≥ 0.80) with a sufficiently large sample size .

The existing version of meta-CART has two shortcomings. Firstly, it is a step-wise procedure. In the first step, the interaction effects are identified by a tree-based algorithm (i.e., CART) using the study effect sizes as outcome variable, and the moderators as predictor variables. In the second step, the moderator effects are tested by a subgroup meta-analysis using the terminal nodes as a new subgrouping variable (with categories referring to the labels of the leaves in which the studies were assigned to by the tree). Secondly, the fixed-effect and random-effects assumptions are not taken into account consistently in meta-CART. The random-effects model assumption is considered by the subgroup meta-analysis in the second step, but not in the splitting procedure of the first step. Furthermore, the fixed-effect model is assumed in the first step, but not in the testing procedure of the second step.

To overcome these shortcomings, we propose two new strategies, one for the fixed effect model and one for the random effects model, that integrate the two steps of meta-CART into one. By applying new splitting criteria and a new splitting algorithm, these new strategies of meta-CART can identify interaction effects and perform the heterogeneity test simultaneously. Furthermore, the model assumption is applied consistently throughout the whole process. The performance of the new strategies of meta-CART are evaluated via an extensive simulation study with different types of moderators (i.e., dichotomous, nominal, ordinal and continuous). The outline of this paper is as follows. First, we describe shortly the fixed-effect and random effects model in meta-analysis. Second, we introduce the new strategies of meta-CART as fixed-effect meta-CART and random-effects meta-CART with an illustrative example using a real-world data set. We then evaluate the performance of the two approaches in a simulation study. Finally, we summarize and discuss the results.

3.2 CART

CART is a recursive partitioning method proposed by Breiman et al. (1984). CART includes two types of trees: classification trees (for a categorical outcome variable) and regression trees (for a continuous outcome variable). In this article, we focus on regression trees for meta-analysis, using a continuous outcome variable (i.e., the study effect size). A previous study showed that in this framework regression trees outperformed classification trees (Li, Dusseldorp, & Meulman, 2017). For a complete introduction for both classification and regression trees, we refer to Merkle and Shaffer (2011).

There are two sequential procedures involved to fit a regression tree: a partitioning procedure that grows a tree to split study cases into more homogeneous subgroups, and a pruning procedure that removes spurious splits from the tree to prevent overfitting. The partitioning procedure starts with all cases in one group (i.e., the root node). Then the

root node is split into two subgroups (i.e., offspring nodes), by searching all possible split points across all predictor variables to find the split that induces the highest decrease in heterogeneity (called impurity). The within-node sum of squares is often used as the impurity for a regression tree. Within a node j, the impurity can be written as

$$i(j) = \sum_{(x_k, d_k) \in j} (d_k - \bar{d}(j))^2, \tag{3.1}$$

where $(x_k, d_k) \in j$ denotes the cases (e.g., studies in meta-analysis) that are assigned to node j with x_k being the predictor vector (e.g., moderators) and d_k being the outcome variable (e.g., the study effect size); $\bar{d}(j)$ is the mean of d_k for all cases (x_k, d_k) that fall into node j (see also Breiman et al., 1984).The partitioning process can be repeated on the offspring nodes, and each split partitions the parent node into two offspring nodes.

For example, in the tree of Figure 3.1b, a predictor variable "T1" with two values "Yes" and "no", which indicates if the behavior change technique "T1: provide information about behavior-health link" was applied in a health psychological intervention, is selected as the first splitting variable. If an intervention has applied "T1", it belongs to the left offspring node. Otherwise, it belongs to the right offspring node. Each of the two offspring nodes can be the candidate of the parent node for the next split.

It is difficult to decide an optimal point to stop the splitting process. Instead, an initial tree is grown as large as possible, and then pruned back to a smaller size by the pruning procedure. To prune a tree, cross-validation is performed to estimate the sum of squared errors.[1] Based on the cross-validation error, there are several pruning rules to select the best size of the tree. To generalize the pruning rules, a pruning parameter c can be introduced to select the pruned tree by using the $c \cdot SE$ rule (Dusseldorp et al., 2010). The $c \cdot SE$ rule selects the smallest tree with a cross-validation error that is within the minimum cross-validation error plus the standard error multiplied by c. For standard CART algorithm, Breiman et al. (1984) suggested using the one-standard-error rule to reduce the instability, which can be regarded as a special case of the $c \cdot SE$ rule when c equals 1.

CART is capable of handling high-dimensional predictor variables of mixed types, and excels in dealing with complex interaction effects. It also has the advantage of straightforward interpretability of the analysis results. However, there are two difficulties when applying standard CART in meta-analysis: (1) the studies are not weighted by their accuracy, and (2) no model assumption is imposed on the algorithm, whereas fix effect and random effects assumptions are used in meta-analysis. We address these two issues and

[1] Ten-fold cross-validation is generally recommended by Breiman et al. (1984). Ten-fold cross-validation involves splitting up one dataset to ten folds. To estimate the cross-validation errors, one fold can be used as the "validation" set, and the left nine folds are used as the "training" set. For each fold used as the validation set, a tree is built using the corresponding training set, and the prediction errors are examined on the validation set.

propose solutions in the following sections.

3.3 Fixed effect and random effects model in subgroup meta-analysis

There are two families of statistical models in meta-analysis: fixed effect (FE) models and random effects (RE) models (Hedges & Olkin, 1985). In this section, we mainly focus on the two models in subgroup analysis, that is, the analysis to evaluate the effect of one categorical moderator in meta-analysis.

Denote the observed effect size of the k^{th} study by d_k, FE models assume that

$$d_k = \delta + \epsilon_k, \tag{3.2}$$

where δ denotes the common effect size for all studies, and ϵ_k is the difference between the observed effect size and the true effect size. There is only one source of variance, the within-study sampling error variance $\sigma^2_{\epsilon_k}$. In FE meta-analysis, the summary effect size is computed as the weighted mean with weights $w_k = 1/\sigma^2_{\epsilon_k}$:

$$d_+ = \frac{\sum d_k/\sigma^2_{\epsilon_k}}{\sum 1/\sigma^2_{\epsilon_k}}. \tag{3.3}$$

In RE models, by contrast, there are two sources of variance: the within-study sampling error variance and the between-studies variance. The observed effect size d_k is assumed to be

$$d_k = \delta + \tau_i + \epsilon_k, \tag{3.4}$$

where δ is the grand mean of population effect sizes, and τ_i is the deviation of the study's true effect size from δ. The summary effect size is computed with weights $w^*_k = 1/(\sigma^2_{\epsilon_k} + \sigma^2_\tau)$:

$$d^*_+ = \frac{\sum d_k/(\sigma^2_{\epsilon_k} + \sigma^2_\tau)}{\sum 1/(\sigma^2_{\epsilon_k} + \sigma^2_\tau)}. \tag{3.5}$$

When study features are available in a meta-analysis, one may perform a subgroup analysis. If a subgroup analysis assumes that the variation of observed effect sizes is only due to the subgroup membership and the within-study sampling error, the FE model is used and it allows for no residual heterogeneity. Under these assumptions, the Q-statistic within the j^{th} subgroup will be

$$Q_j = \sum_{k=1}^{K_j} \frac{(d_{jk} - d_{j+})^2)}{\sigma^2_{\epsilon_{jk}}}, \tag{3.6}$$

where K_j is the number of studies in the j^{th} subgroup, d_{jk} is the observed effect size of the the k^{th} study in the j^{th} subgroup, and d_{j+} is the subgroup weighted mean.

The between-subgroups Q-statistic is given by

$$Q_B = \sum_{j=1}^{J} \sum_{k=1}^{K_j} \frac{(d_{j+} - d_{++})^2}{\sigma^2_{\epsilon_{jk}}}, \tag{3.7}$$

where J is the total number of subgroups, and d_{++} is the grand weighted mean.

The total weighted sum of squares for all studies is

$$Q_T = \sum_{j=1}^{J} \sum_{k=1}^{K} \frac{(d_{jk} - d_{++})^2)}{\sigma^2_{\epsilon_{jk}}}. \tag{3.8}$$

There is a simple relationship among Q_j, Q_B and Q_T that is analogous to the partitioning of the sum of squares in analysis of variance (Hedges & Olkin, 1985, Page 156),

$$Q_T = \sum_{j=1}^{J} Q_j + Q_B. \tag{3.9}$$

If a subgroup analysis assumes that residual heterogeneity exists, the RE model is used and it allows for variation unexplained by the subgroup membership and the within-study sampling error. For subgroup analysis using a RE model, a generally advocated approach is to assume a fixed effect model across subgroups and a random effects model within subgroups [2] (Borenstein et al., 2009). This assumption means that the variation in subgroup means is only explained by the subgroup membership, and the variation in the observed study effect sizes is due to the subgroup membership, the residual heterogeneity between studies, and the within-study sampling errors. The residual heterogeneity can be estimated separately within subgroups, or a common estimate to all studies can be computed by pooling the within-subgroup estimates. There are several estimators for the residual heterogeneity available. In this study, we compute the pooled estimate for residual heterogeneity using the DerSimonian and Laird (1986) method. The pooled residual heterogeneity is computed as

$$\sigma^2_\tau = \frac{\sum_{j=1}^{p} Q_j - \sum_{j=1}^{p} df_j}{\sum_{j=1}^{p} C_j}, \tag{3.10}$$

where Q_j is computed as in (3.6), df_j equals $K - 1$, and the components C_j using the fixed effects weights, are computed as

[2] sometimes called mixed effects model.

$$C_j = \sum_{k=1}^{K} w_{jk} - \frac{\sum w_{jk}^2}{\sum w_{jk}}. \tag{3.11}$$

The between-subgroups Q-statistic is given by

$$Q_B^* = Q_T^* - \sum_{j=1}^{p} Q_j^*, \tag{3.12}$$

where

$$Q_T^* = \sum_{j=1}^{p} \sum_{k=1}^{K} \frac{(d_{jk} - d_{++}^*)^2}{\sigma_{\epsilon_{jk}}^2 + \sigma_\tau^2}, \tag{3.13}$$

and

$$Q_j^* = \sum_{k=1}^{K} \frac{(d_{jk} - d_{j+}^*)^2}{\sigma_{\epsilon_{jk}}^2 + \sigma_\tau^2}. \tag{3.14}$$

3.4 Fixed effect meta-CART

3.4.1 The algorithm

To solve the two difficulties when applying standard CART in meta-analysis, FE meta-CART applies weights in the CART algorithm, and assumes absence of residual heterogeneity when searching for the influential moderators. In FE meta-CART, we apply the weights used in the FE models in meta-analysis ($w_k = 1/\sigma_{\epsilon_k}^2$). As a result, the weighted within-node sum of squares will be equivalent to the Q-statistic within node j. Denote the weighted mean of the outcome variable in node j as $d_+(j)$. It can be shown that

$$d_+(j) = \frac{\sum_{(x_k,d_k)\in j}(d_k \cdot w_k)}{\sum_{(x_k,d_k)\in j}(w_k)} = \frac{\sum_{(x_k,d_k)\in j} d_k/\sigma_{\epsilon_k}^2}{\sum_{(x_k,d_k)\in j} 1/\sigma_{\epsilon_k}^2}, \tag{3.15}$$

which is equal to the summary effect size in node j under the fixed effect assumption (see [3.3]). Also, the impurity function can be computed as

$$i(j) = \sum_{(x_k,d_k)\in j} w_k(d_k - d_+(j))^2 = \sum_{(x_k,d_k)\in j} \frac{(d_k - d_+(j))^2}{\sigma_{\epsilon_k}^2}, \tag{3.16}$$

which is equal to the Q-statistic within node j as in (3.6).

When growing a FE meta-regression tree, the algorithm searches for the moderator and the split point that minimize the sum of Q_j of the offspring nodes. Note that this is equal to the split that maximizes Q_B (see [3.9]). The splitting process continues until all terminal nodes contain only one or two studies. Then the initial tree will be pruned to a smaller size using cross-validation to prevent overfitting. For the previous version of meta-CART, a pruning rule with $c = 0.5$ was generally recommended by Li, Dusseldorp,

and Meulman (2017). For the new strategies of meta-CART in this study, we apply two pruning rules with $c = 0.5$ and $c = 1$ and examine their performance. After the pruning process, the final tree gives the corresponding between-subgroups Q_B and the estimates for summary effect sizes d_{j+} within each subgroup as the analysis results.

3.4.2 An illustrative example

To illustrate the algorithm, we will use the data from Michie, Abraham, et al. (2009) as an example. The complete data consist of 101 studies reporting 122 interventions targeted at physical activity and healthy eating. In this motivating example, we will re-analyze these data focusing on the motivation-enhancing behavior change techniques (BCTs) that may explain the heterogeneity in the effect sizes of interventions. The interventions that include at least one of the motivation-enhancing BCTs were selected ($N = 106$). The details about the motivation-enhancing BCTs can be found in Table 3.1.

To identify influential BCTs and the interaction effects between them, FE meta-CART starts with a root node including all selected studies (Figure 3.1a). For the first split, the algorithm selects the moderator T1 since it results into the largest between-subgroups Q-statistic ($Q_B = 17.19$ among 0.004, 0.10 and 4.35 when choosing the splitting variable as T2, T4 and T5, respectively.) The root node is thereby split into two children nodes (Figure 3.1b). These two nodes then become the candidates for the parent node for the second split. The algorithm searches through all the combinations of parent node and splitting variable, and selects the combination that maximizes the Q_B. This splitting process continues until a large tree is grown and all of the terminal nodes only contain one or two studies. Then the large tree is pruned to a smaller size by the cross-validation procedure, and the final tree is selected as a tree with three terminal nodes shown in Figure 3.1c.

The final tree represents an interaction effect between the BCTs "T1: provide information about behavior-health link" and "T4: prompt intention formation". The main result of this tree is that the combination of "T1" and "T4" results in the highest effect size. More specifically, when "T1" is not applied, the average effect size of the interventions is 0.20 . When "T1" is applied together with "T4", the interventions have the highest average effect size (0.44). When only "T1" is applied without "T4", the average effect size is 0.19. The estimated subgroup effect sizes and the between-subgroups Q-statistic (Q_B) are obtained simultaneously as the tree is grown. The final fixed effect Q_B equal to 40.59 indicates a significant interaction effect (p-value <0.001, $df = 2$).

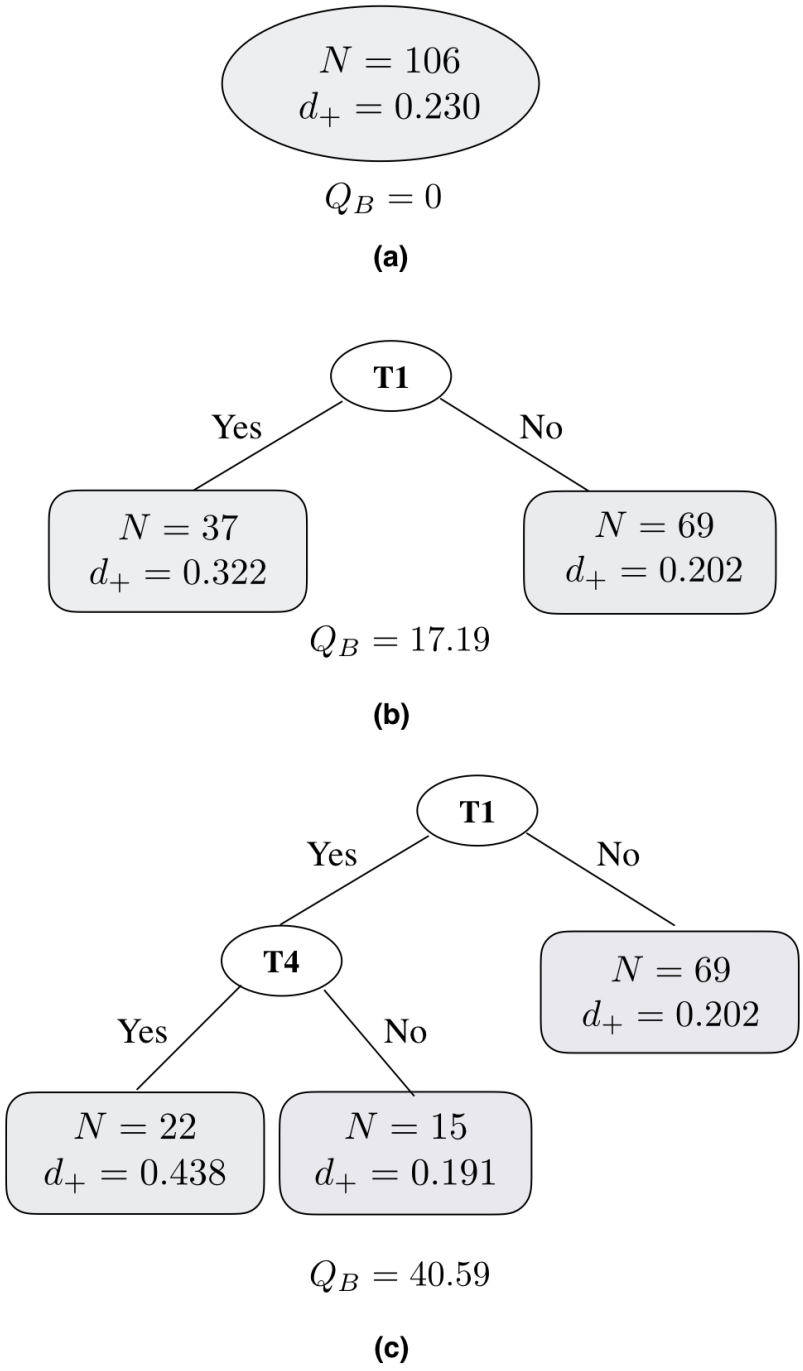


Figure 3.1: The first three splits of a FE meta-tree for the studies that applied at least one of the motivation-enhancing techniques in Michie, Abraham, et al. (2009). T1 and T4 are labels for behavior change techniques "Provide information about behavior-health link" and "Prompt intention formation", respectively.

Table 3.1: overview of the motivation-enhancing behavior change techniques. The last column displays the number (#) of studies that applied a technique in Michie, Abraham, et al. (2009).

Technique	Definition	#
1. Provide information about behavior-health link	General information about behavior risk, for example, susceptibility to poor health outcomes or mortality risk in relation to the behavior	37
2. Provide information on consequences	Information about the benefits and costs of action or inaction, focusing on what will happen if the person does or does not perform the behavior	64
3. Provide information about other's approval	Information about what others think about the person's behavior and whether others will approve or disapprove of any proposed behavior change	0
4. Prompt intention formation	Encouraging the person to decide to act or set a general goal, for example, to make a behavior resolution, such as "I will take more exercise next week"	74
5. Motivational interviewing	Prompting the person to provide self-motivating statements and evaluations of their own behavior to minimize resistance to change	17

3.5 Random effects meta-CART

3.5.1 The algorithm

RE meta-CART takes residual heterogeneity into account, and searches for the influential moderators based on the RE between-subgroups Q-statistic (Q_B^*) as given in (3.12). To grow a RE meta-tree, the algorithm starts with a root node that consists of all studies. In each split of the algorithm, all terminal nodes of the tree obtained from the previous step are considered as candidate parent nodes. To choose a split, two substeps are performed. The first substep is to examine in each candidate parent node the optimal combination of a splitting moderator variable and a split point. By each possible combination of the splitting variable and split point, the candidate parent node can be split into two offspring nodes and a new branch is formed after the split. For this split, the residual heterogeneity unexplained by the subgroup membership is estimated for the whole tree and the corresponding Q_B^* is computed. The first substep then is concluded by selecting across all possible splits the optimal combination that maximizes the Q_B^*. In the second substep, the values of Q_B^* associated with the optimal combination are compared across all candidate parent nodes, and the node with the highest Q_B^* will be chosen. After these

two substeps, a split is made by splitting the chosen parent node into two offspring nodes (on the basis of the optimal combination of the splitting variable and the split point associated with that parent node).

Same as the splitting process in FE meta-CART, each new split in RE meta-CART refreshes the partitioning criterion: the between-subgroups Q-statistic. However, the RE model implies that the residual heterogeneity σ_τ^2 is re-estimated after each split. As a result, a split within one node will globally affect the estimation of σ_τ^2 and the value of Q_B^*. In other words, the within-subgroup Q_j^* needs be computed not only for the new offspring nodes, but also for all the other existing terminal nodes in the current tree. Thus, this partitioning method is not fully recursive. Instead, RE meta-CART applies a sequential partitioning algorithm.

The pruning process of RE meta-CART is the same as FE meta-CART. The initial large tree is pruned back to a smaller size using cross-validation with the $c \cdot SE$ rule. The associated between-subgroups Q_B^*, the estimates for residual heterogeneity σ_τ^2, and the within-subgroup summary effect sizes d_{j+}^* are obtained as the final tree is selected.

3.5.2 An illustrative example

We will use the same data as in 3.4.2 to illustrate the RE meta-CART algorithm. To identify the interaction effects using the random effects model, the algorithm starts with a root node including all selected studies (Figure 3.2a). The first split selects the moderator T1, which results into the largest between-subgroups Q-statistic ($Q_B^* = 2.74$ among 0.24, 1.32 and 0.10 when choosing the splitting variable as T2, T4 and T5, respectively.) Then the two children nodes as shown in Figure 3.2b become the candidates of the parent node for the second split. For the second split, the algorithm searches through all the combinations of parent node and splitting variable, and selects the combination that maximizes the Q_B^*. Note that the value of the summary effect size in the unselected node d_{1+}^* has been slightly changed from 0.245 to 0.241 after the new split. This change is due to the new estimate for the residual heterogeneity σ_τ^2. Therefore, a new split influences not only the selected parent node but also the unselected node(s). As a result, the sequence of the splits globally influences the estimates for σ_τ^2, d_+^* and Q_B^*. This sequential partitioning process continues until a large tree is grown and all of the terminal nodes only contain one or two studies[3]. After the pruning process, the final tree is selected as a tree with three terminal nodes shown in Figure 3.2c.

The final tree by RE meta-CART selects the same moderators as FE meta-CART in 3.4.2: "T1: provide information about behavior-health link" and "T4: prompt intention formation". But under the RE assumption, the estimated summary effect sizes in each subgroup and the between-subgroups Q-statistic are different from those estimated using

[3]The exact minimal number of studies in a node is fixed before splitting. We used here a size of two.

FE model. The random effects $Q_B^* = 13.20$ indicated a significant interaction effect (p-value $= 0.001$, $df = 2$).

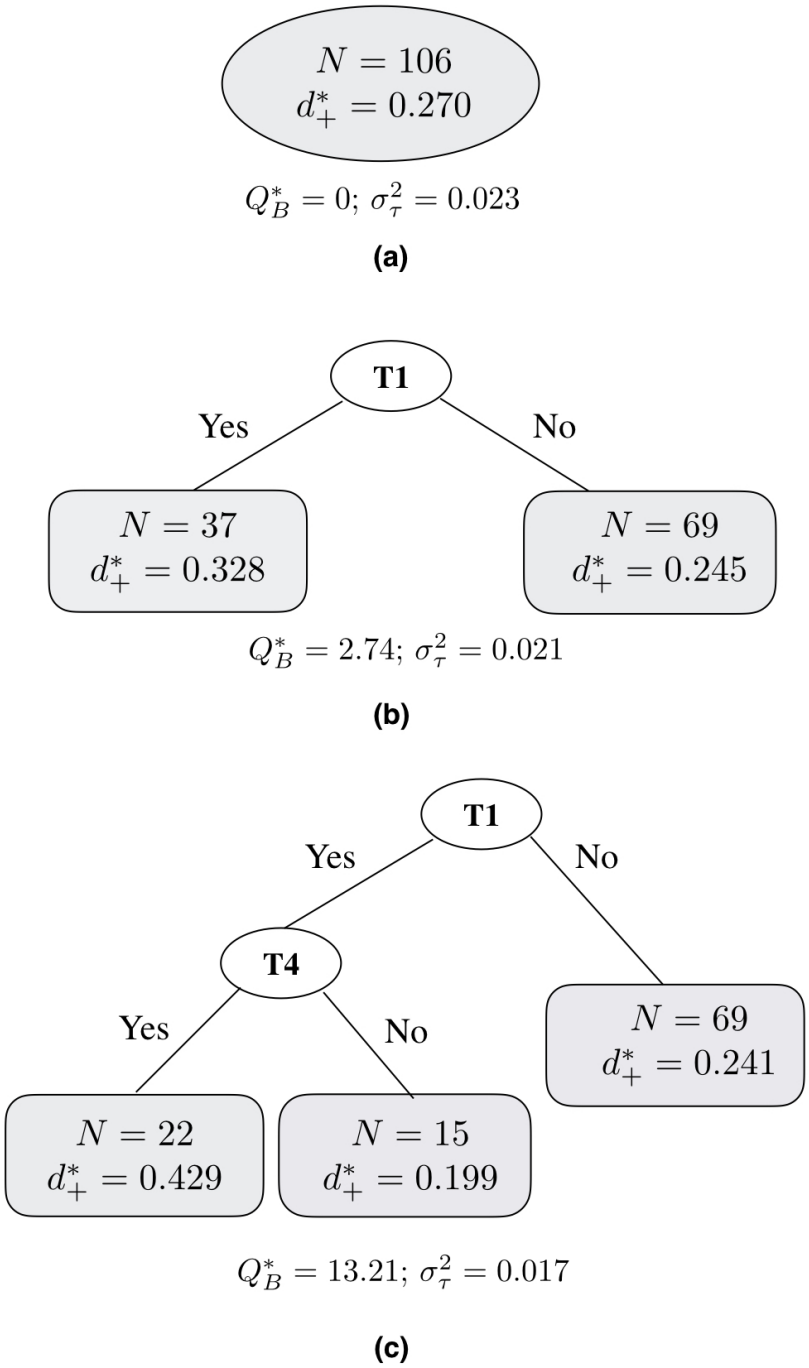


Figure 3.2: The first three splits of a RE meta-tree for the studies that applied at least one of the motivation-enhancing techniques in Michie, Abraham, et al. (2009). T1 and T4 are labels for behavior change techniques "Provide information about behavior-health link" and "Prompt intention formation", respectively.

3.6 Simulation

3.6.1 Motivation

In the simulation study, we first aim at selecting pruning rules for the new strategies of meta-CART to control the risk of finding spurious effects (see 3.4.1). Secondly, we evaluate the performance of FE meta-CART and RE meta-CART under various conditions using the selected pruning rules. It is important to note that the simulation study does not aim at comparing FE meta-CART and RE meta-CART. The choice of the model assumption should be based on theoretical grounds (also see 3.8.2). The conditions that we consider include observable features of meta-analytic data sets, such as the number of studies, the within-study sample sizes, the type of moderators, and the number of moderators, as well as unobservable structures and parameters underlying the data, such as the complexity of the interaction effects, the magnitude of the interaction effect, the correlation between moderators, and the residual heterogeneity. The recovery performance of meta-CART is measured by the ability of successfully retrieving the true structures underlying the data. In addition, we compare its performance to meta-regression with true structures specified beforehand, which can be seen as an idealized solution.

We use a design for the true tree structures that is comparable to Li, Dusseldorp, and Meulman (2017). Five tree structures with increasing complexity are used as the underlying true model to generate data sets (see Figure 3.3). Model A was used to assess the probability that meta-CART falsely identifies (a) moderator effect(s) when there is no moderator in the true model (Type I error). Model B was used to evaluate the ability of meta-CART to identify the main effect of a single moderator. Models C, D and E were used to evaluate the extent to which meta-CART correctly identifies the interaction effects between moderators when interaction effects are present in the true model. In the designed tree model, the treatment is effective only in studies with certain combination(s) of study features. The studies are thereby split on moderators into subgroups. The average effect size in the ineffective subgroups is fixed to be 0, and the average effect size in the effective subgroups was a design factor and is denoted by δ_I. The true effect sizes of the studies are generated from a normal distribution with mean equal to the average effect size (i.e., 0 for ineffective subgroups and δ_I for effective subgroups), and standard deviation equal to the residual heterogeneity.

3.6.2 Design factors

Artificial data were generated with observed study effect sizes d, the within-study sample size n and potential moderators $x_1, \ldots, x_M$. We used three design factors concerning the moderators. The total number of potential moderators M was a design factor with three values: 5, 10 and 20. We generated four different types of moderator variables ($Type$):

binary, *nominal*, *ordinal*, and *continuous*. In our study, all the ordinal moderators and nominal moderators were generated with three levels (1, 2, 3 for ordinal and A, B, C for nominal). The correlation matrix between the moderators ($\mathbf{R}$) was a design factor. Both independent and correlated moderators were generated. To generate uncorrelated moderators we use $\mathbf{R} = \mathbf{I}$ as the population correlation matrix. To generate correlated moderators, we used a correlation matrix $\mathbf{R}$ computed from a real-world data set by Michie, Abraham, et al. (2009). We first randomly sample M moderators from the 26 moderators in the Michie, Abraham, et al. (2009) data and compute the correlation matrix. Then we generate M moderators using the computed correlation matrix. The range of correlations varies roughly between -0.40 and 0.40.

In addition to the three design factors concerning the moderators, four other design factors that may influence the effect size d were examined: (a) the number of studies (K); (b) the average within-study sample size ($\bar{n}$); (c) the residual heterogeneity (σ_τ^2); and (d) the magnitude of the interaction effect (δ_I). Three values of K were chosen: 40, 80, 120. Because a previous study showed that meta-CART applied to data sets with $K \leq 20$ studies results in poor power rates (≤ 0.30) (Li, Dusseldorp, & Meulman, 2017), therefore we start with $K = 40$. We used the same method as in Viechtbauer (2007b) to generate the within-study sample size n_k; the values of n_k were sampled from a normal distribution with an average sample size $\bar{n}$ and standard deviation $\bar{n}/3$. Three levels of the average within-study sample size $\bar{n}$ were chosen as 40, 80, 160. The resulting n_k ranged roughly between 15 and 420, which are plausible values encountered in practice. The values of the residual heterogeneity unexplained by the moderators σ_τ^2 were chosen as 0, 0.025, 0.05. The values of δ_I were chosen as 0.3, 0.4, 0.5 and 0.8, among which 0.5 and 0.8 corresponding to a medium and a large effect size, respectively (Cohen, 1988). A small effect size $\delta_I = 0.2$ was not included in the study, because the previous study showed that meta-CART failed to have enough power to detect small interaction effect(s) (Li, Dusseldorp, & Meulman, 2017). Thus in total we have $M \times Type \times \mathbf{R} \times K \times \bar{n} \times \sigma_\tau^2 \times \delta_I = 3 \times 4 \times 2 \times 3 \times 3 \times 3 \times 4 = 2592$ design factors.

3.6.3 Monte Carlo Simulation

For each of the five tree structures, 1000 data sets were generated with all possible combinations of design factors (i.e., $2592 \times 5 \times 1000 = 12,960,000$ data sets). To generate continuous moderators, we first generate continuous variables from a multivariate normal distribution with variable means equal to 20, standard deviations equal to 10, and with a correlation matrix as identity matrix (for independent moderators) or a correlation matrix computed as mentioned above (for correlated moderators). Then the generated variables were rounded to the first decimal place to allow for duplicate values. The average number of unique values of the continuous moderators was 37, 71, 102 for $K = 40$, 80,

120, respectively. For non-continuous moderators, we first randomly generate continuous variables from a multivariate normal distribution with a correlation matrix as mentioned above. For binary moderators, the generated continuous variables were dichotomized around their mean. For nominal moderators, the continuous variables were split by the 1/3 quantile and 2/3 quantile of the normal distribution, and the resulting three intervals were randomly labeled by the letters A, B and C. For ordinal moderators, the continuous variables were split by the 1/3 quantile and 2/3 quantile of the normal distribution and ordered by the intervals that they belonged to. Note that the polytomization attenuates the correlations between the resulting variables.

With the given moderators and the tree structure, the average true effect size Δ_j was computed for each subgroup j. For a single study k within each subgroup j, the true effect size δ_{jk} was sampled from a normal distribution with mean Δ_j and variance σ_τ^2. Finally, the observed effect size d_{jk} was sampled from a non-central t-distribution, and the corresponding sampling errors σ_ϵ^2 were calculated (see Appendix A.2 for detailed information).

3.6.4 The evaluation criteria for success

Three criteria are used to judge the performance of meta-CART with respect to the true model underlying the data:

Criterion 1. Meta-CART falsely detects the presence of moderator effect(s) in the data sets generated from model A (Type I error).

Criterion 2. Meta-CART detects the presence of moderator effect(s) in the data sets generated from model B, C, D or E (power). This criterion evaluates if a non-trivial tree is detected (i.e., a pruned tree with at least one split and a significant between-subgroups Q), but does not examine the size of the tree and the correct moderator(s).

Criterion 3. Meta-CART successfully selects the moderators used in the true model (recovery of moderator(s)). This criterion examines if the true model is fully recovered with all the true moderators and no spurious moderators are selected.

The computation of these criteria will be specified in section 3.6.7.

3.6.5 Comparison to meta-regression

FE and RE meta-regression analyses were performed on the datasets generated from non-trivial trees (models B, C, D and E) with the true moderator effect(s) specified. The analyses results were compared to meta-CART in terms of recovery of moderators (criterion 3). Note that in this scenario meta-regression is expected to result in higher

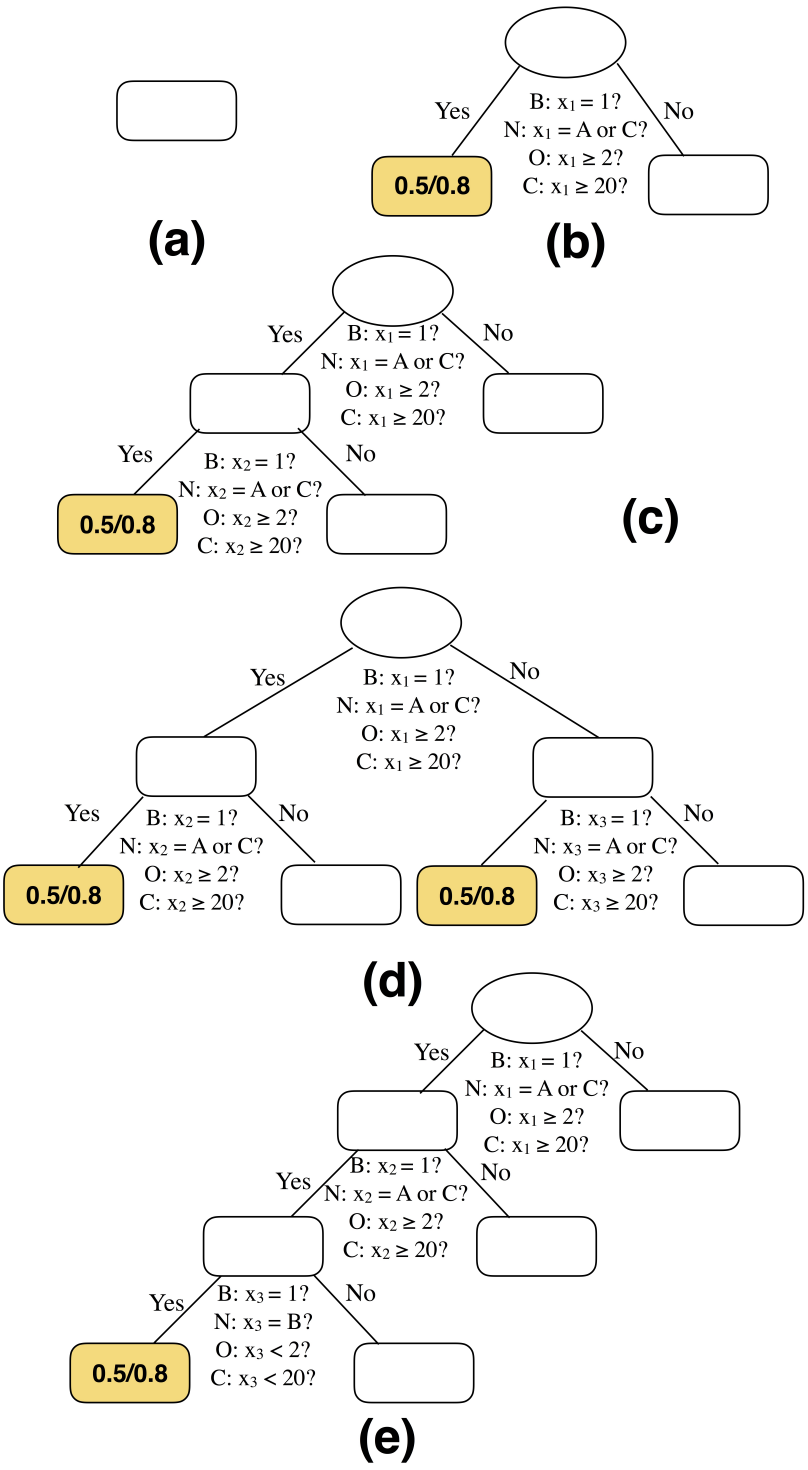


Figure 3.3: Simulated data sets were generated from five true tree structures: (a) to (e). The letters B, N, O and C denote the four types of moderator variables: binary, nominal, ordinal and continuous respectively. These tree structures represents a true model including: no moderator effect (model A); only main effect of one moderator (model B); one two-way interaction (model C); two two-way interactions (model D); and one three-way interaction (model E), respectively.

recovery rates, since the true structure is specified in meta-regression but to be explored by meta-CART. The goal of this comparison is to evaluate how meta-CART compares to the optimal performance that meta-regression can reach in an idealized scenario.

3.6.6 The estimates for subgroup effect sizes

The estimates for subgroup effect sizes were examined in the terminal nodes from the successfully retrieved trees for one cell of the design with medium level of each design factors (i.e., Tree complexity = model C, $K = 80$, $\bar{n} = 80$, $\sigma_\tau^2 = 0.025$, $M = 10$, $\delta_I = 0.5$, $\mathbf{R}$ = the computed correlation matrix, Variable type = ordinal moderators). Although trees with the first splitting variable as x_2 and the second splitting variable as x_1 are also equivalent to model C, only the trees exactly the same as model C shown in Figure 3.3 were examined to make the resulting subgroups comparable. For each terminal node in the selected trees, the averaged subgroup effect size estimates were computed, and the proportion that the 95% confidence intervals (CIs) contain the true value were counted.

3.6.7 Analysis

FE meta-CART and RE meta-CART were applied to each generated data set using two pruning rules with $c = 0.5$ and $c = 1.0$. The significance of the subgrouping defined by the pruned tree was tested by the between subgroups Q-statistic with $\alpha = 0.05$.

In total, $12,960 \times 1000$ analyses were performed per strategy per pruning rule. Each of the three criteria was evaluated and coded with 0 for "not satisfied" and 1 for "satisfied" for each data set. Subsequently, for each cell of the design, the proportion of "satisfied" solutions was computed per criterion. The resulting proportions were subjected to analyses of variance (ANOVA) with the design factors and their interactions as independent variables. Due to the computation time and the difficulty of interpretation, only four-way and lower-order interactions were included as independent variables, and the higher-order interactions were used as error terms. Partial eta squared ($\hat{\eta}_P^2$, see Keppel, 1991, pp. 222-224) was computed for all main effects and interaction terms. For Type I error rate, the pruning parameter c was included as a within-subject design factor, and the generalized eta squared ($\hat{\eta}_G^2$, see Olejnik & Algina, 2003) was computed for all main effects and interaction effect terms.

Both FE and RE meta-CART were compared to meta-regression on the $9,720 \times 1000$ data sets generated from non-trivial trees. For meta-regression, criterion 3 is defined as all the true moderator effects being significant (i.e., p-value < 0.05). For each cell of the design, the proportion of this criterion being satisfied was computed as the recovery rate. The difference in recovery rates between meta-CART and meta-regression within each cell were subjected to analyses of variance (ANOVA) as mentioned above.

The simulation, the meta-CART analyses, the meta-regression analyses, and ANOVA

were performed in the R language (Team, 2017). The meta-CART analyses were performed using the R-package **metacart** (Li, Dusseldorp, Liu, & Meulman, 2017). The meta-regression analyses were performed using the R-package **metafor** (Viechtbauer, 2010). The R-codes for the simulation study are available at `https://osf.io/mghsz/`.

3.7 Results

For the Type I error rate of FE meta-CART, the ANOVA results reveal that the number of studies (K) and the pruning parameter c have much stronger influence than the other design factors (see Supporting Material Table 3.5). For the Type I error rate of RE meta-CART, the main effect of K and the interaction between K and c have the strongest influence (see Table 3.6). For both FE meta-CART and RE meta-CART, the estimated Type I error rates decrease with increasing K and the pruning parameter c. Table 6.2 shows the estimated Type I error rates averaged over the less influential design factors (i.e., $Type$, $\mathbf{R}$, M, $\bar{n}$, σ^2_τ, δ_I). An average Type I error below .05 was chosen to be acceptable in order to control for the risk of finding spurious (interaction) effects. Therefore, the best pruning parameter for FE meta-CART was selected as $c = 1$ if $K < 80$ and $c = 0.5$ if $K \geq 80$. And the best pruning parameter for RE meta-CART was selected as $c = 1$ if $K < 120$ and $c = 0.5$ if $K = 120$. A higher value of c indicates more pruning. Thus, for smaller K a higher amount of pruning is needed to control Type I error.

For the power rates and the recovery rates of the moderators, ANOVA was employed to analyze the results of meta-CART using the selected pruning parameters as defined above. For power rates, the ANOVA results on recovery rates reveal that FE meta-CART and RE meta-CART are both strongly influenced by the main effects of the number of studies K, the magnitude of the interaction effect δ_I, the tree complexity (B, C, D or E), and the residual heterogeneity σ^2_τ (Tables 3.7 and 3.8). In addition, RE meta-CART is also strongly influenced by the main effects of the average within-study sample size $\bar{n}$ and the type of moderator variables. Similarly, the recovery rates of FE meta-CART and RE meta-CART are both strongly influenced by the main effects of K, δ_I, the tree complexity, σ^2_τ, and the type of moderator variables (Table 3.9 and 3.10). The recovery rates of RE meta-CART are also strongly influenced by $\bar{n}$. Because the patterns of power

Table 3.2: Type I error rate of meta-CART, averaged over $Type$, $\mathbf{R}$, M, $\bar{n}$, σ^2_τ, δ_I.

		Fixed effect meta-CART			Random effects meta-CART		
model	c	$K = 40$	$K = 80$	$K = 120$	$K = 40$	$K = 80$	K=120
A	0.5	**.071 (.011)**	.037 (.007)	.023 (.006)	**.095 (.023)**	**.061 (.018)**	.042 (.014)
	1.0	.034 (.009)	.010 (.005)	.004 (.003)	.033 (.011)	.012 (.005)	.005 (.003)

[4]Type I error rates higher than 0.05 are in boldface. The numbers in parentheses display the standard deviations of the Type I error rates.

and recovery rates are similar and the latter is the more stringent criterion, we focus on the results concerning recovery rates.

In general, the recovery rates increase with increasing K, δ_I and $\bar{n}$, and decrease with increasing σ_τ^2 and tree complexity. Binary moderators have the highest recovery rates, whereas continuous moderators have the lowest recovery rates. The recovery rates for nominal and ordinal moderators are similar. The influence of K, δ_I, the type of moderator variables and the tree complexity are shown in Figures 3.4, 3.5 and 3.6. When $K = 120$ (see Figure 3.4), both FE and RE meta-CART are able to achieve satisfactory recovery rates (≥ 0.80) for simple moderator effects (models B and C) in most cases, only with some exceptions when $\delta_I = 0.3$ for non-continuous moderators or $\delta_I \leq 0.4$ for continuous moderators. For complex interaction effects (models D and E), meta-CART is able to achieve satisfactory recovery rates if the interaction effect size is large ($\delta_I = 0.8$) depending on the type of moderators. When the moderators are binary variables, meta-CART can always achieve satisfactory recovery rates for $\delta_I \geq 0.8$. When the moderators are nominal or ordinal, FE meta-CART can achieve satisfactory recovery rates for model D, whereas RE meta-CART can achieve satisfactory recovery rates for model E. When the moderators are continuous variables, FE meta-CART can achieve satisfactory recovery rates for model D, but RE meta-CART fails to achieve recovery rates higher than 0.80. When $K = 80$ (see Figure 3.5), both FE and RE meta-CART achieve satisfactory recovery rates for simple moderator effects in most cases, with some exceptions when $\delta_I = 0.3$ for non-continuous moderators or $\delta_I \leq 0.5$ for continuous moderators. For complex interaction effects, both FE and RE meta-CART are able to achieve satisfactory recovery rates for binary moderators if the effect size is large ($\delta_I = 0.8$), but fail to achieve recovery rates higher than 0.80 for non-binary moderators. When $K = 40$ (see Figure 3.6), both FE and RE meta-CART are able to achieve satisfactory recovery rates for simple moderator effects, but fails to achieve recovery rates higher than 0.80 for complex interaction effects. When there is only a univariate moderator effect in the true model (model B), both FE and RE meta-CART have good performance in most cases. When there is a two-way interaction (model C), both FE and RE meta-CART are able to achieve satisfactory recovery rates if the moderators are non-continuous and the interaction effect size is large ($\delta_I = 0.8$).

For both FE and RE model assumption, the ANOVA results reveal that the difference in recovery rates between meta-CART and meta-regression are strongly influenced by the tree complexity (Tables 3.11 and 3.12). Table 3.3 shows the recovery rates and the difference averaged over the less influential design factors (i.e., K, $Type$, $\mathbf{R}$, M, $\bar{n}$, σ_τ^2, δ_I). For simple moderator effects, the recovery rates of meta-CART are close to meta-regression with the correct structure specified. For complex interaction effects, the difference is larger.

From the 1000 data sets generated from the cell of the design described in 3.6.6, 461

trees were selected to examine the estimates for the subgroup effect sizes and the coverage of 95% CIs. Table 3.4 shows that the averaged estimates are close to the true values in both ineffective subgroups ($\delta = 0$) and effective subgroups ($\delta = 0.5$). The 95% CIs of FE meta-CART have lower coverage than the nominated coverage probability, whereas the 95% CIs of RE meta-CART have coverage close to 0.95 for all three subgroups.

Table 3.3: Difference in recovery rates between meta-CART and meta-regression, averaged over K, $Type$, $\mathbf{R}$, M, $\bar{n}$, σ^2_τ, δ_I.

	Fixed effect			Random effects		
Tree	meta-CART	meta-regression	difference	meta-CART	meta-regression	difference
B	0.94 (0.13)	1.00 (0.02)	-0.05 (0.11)	0.95 (0.12)	0.99 (0.05)	-0.04 (0.08)
C	0.71 (0.32)	0.83 (0.19)	-0.12 (0.20)	0.71 (0.33)	0.71 (0.27)	-0.00 (0.19)
D	0.33 (0.35)	0.72 (0.28)	-0.39 (0.24)	0.22 (0.27)	0.56 (0.34)	-0.34 (0.20)
E	0.21 (0.28)	0.74 (0.27)	-0.53 (0.24)	0.24 (0.31)	0.60 (0.32)	-0.35 (0.26)

[5]The difference is computed as the averaged recovery rates of meta-CART subtracted by the averaged recovery rates of meta-regression. The numbers in parentheses display the standard deviations.

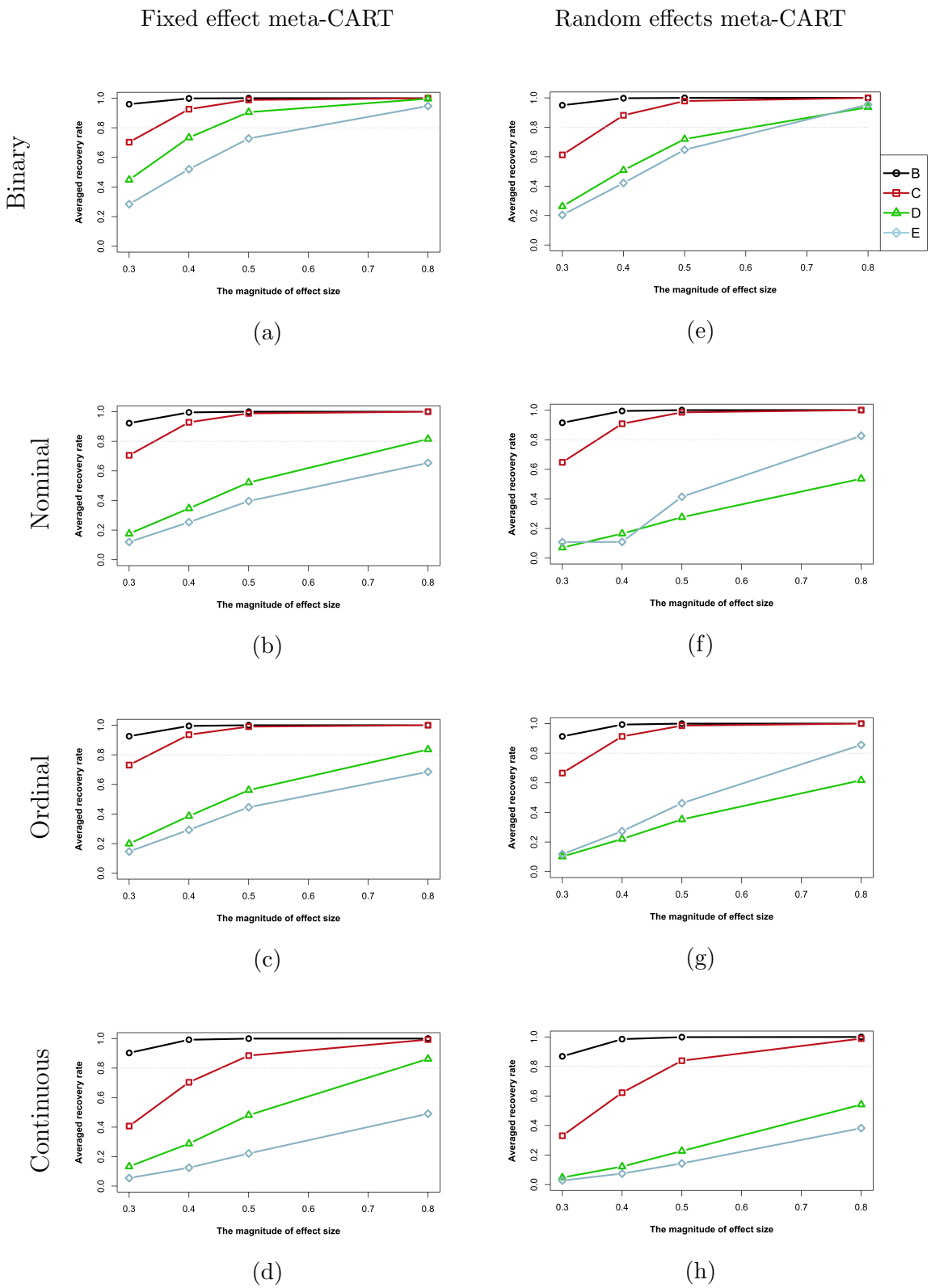


Figure 3.4: Recovery rates of FE and RE meta-CART when $K = 120$. Separate plots are shown for the types of moderator variables. Separate lines are shown for the tree complexity (models B, C, D and E).

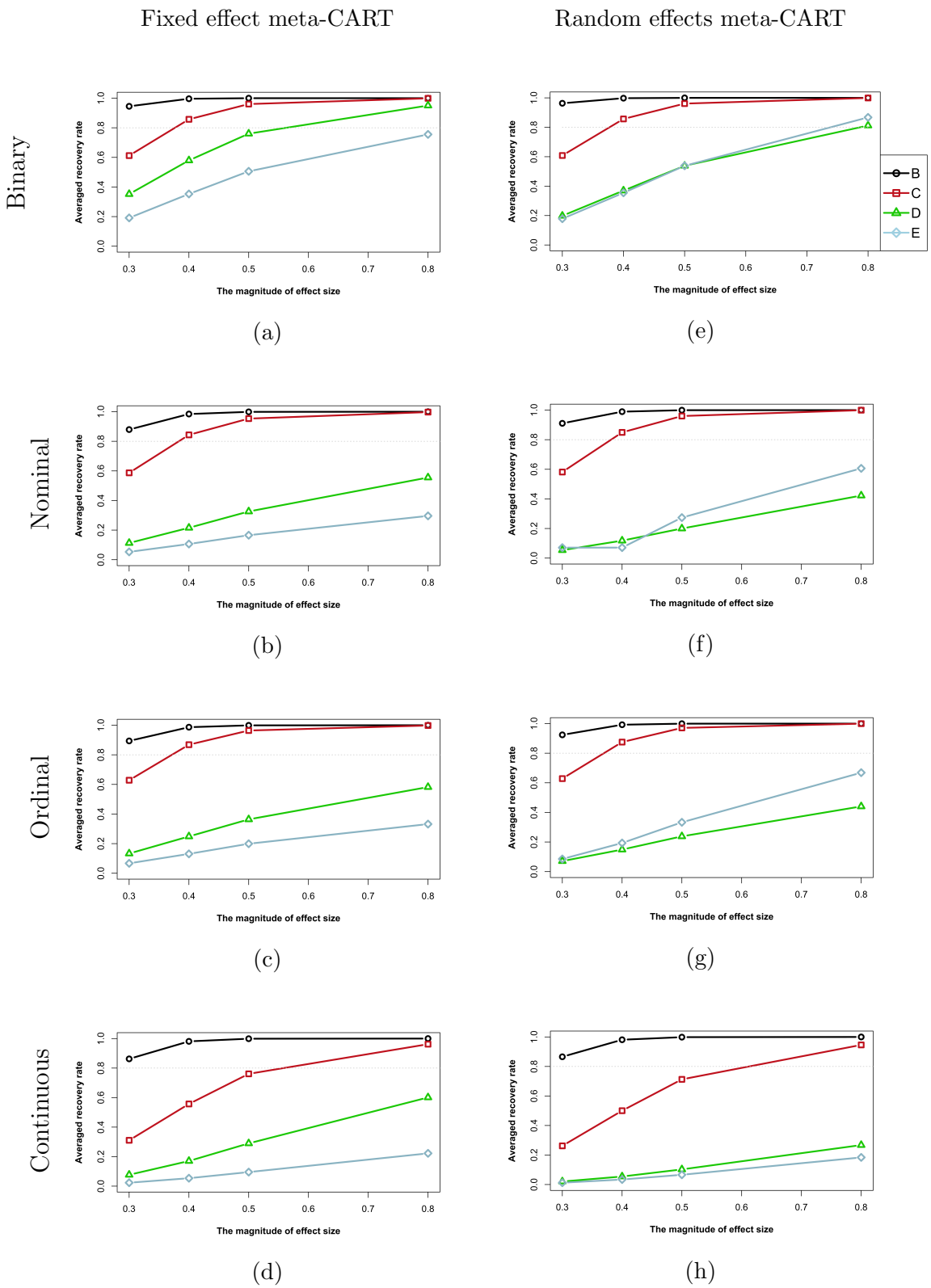


Figure 3.5: Recovery rates of FE and RE meta-CART when $K = 80$. Separate plots are shown for the types of moderator variables. Separate lines are shown for the tree complexity (models B, C, D and E).

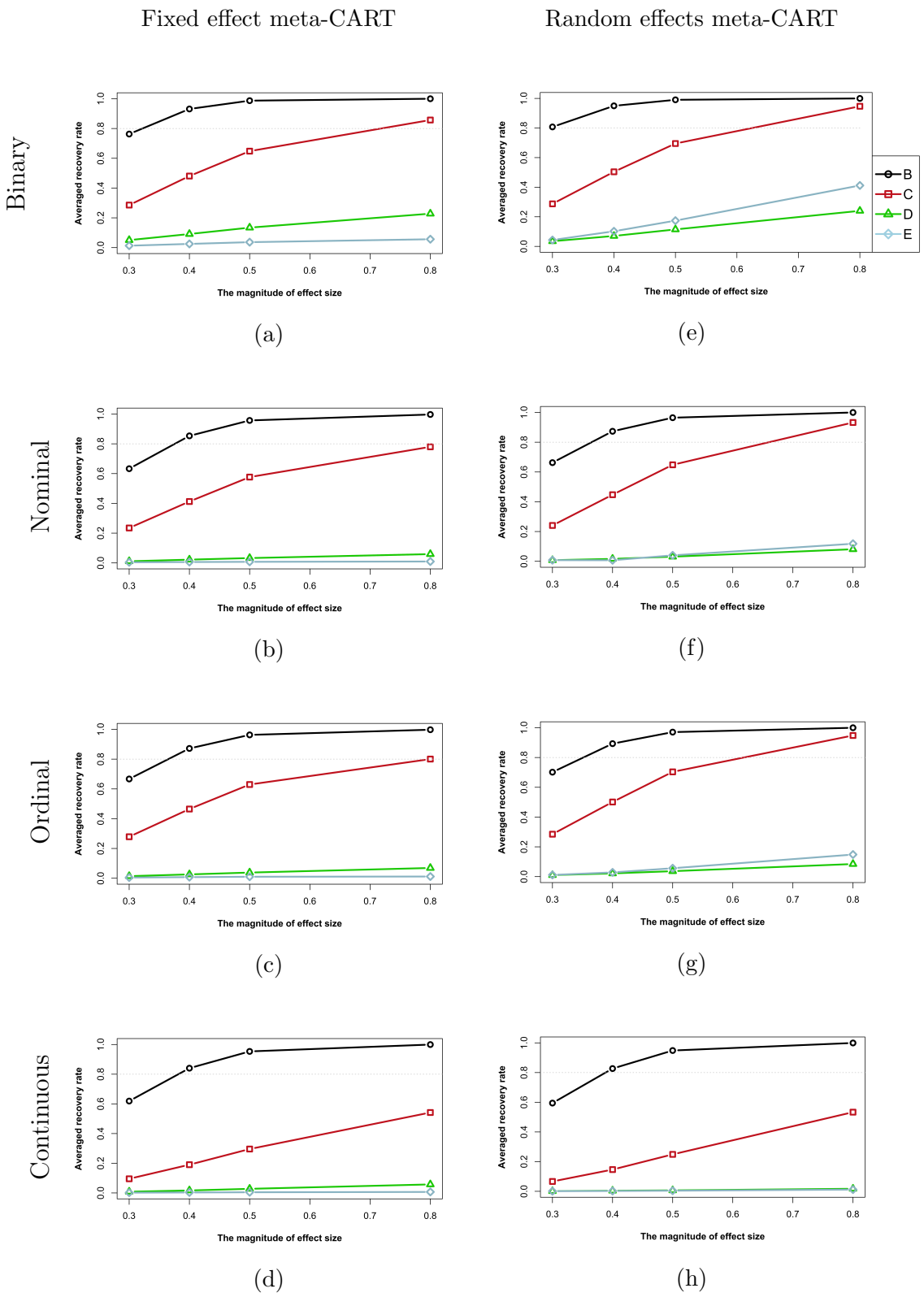


Figure 3.6: Recovery rates of FE and RE meta-CART when $K = 40$. Separate plots are shown for the types of moderator variables. Separate lines are shown for the tree complexity (models B, C, D and E).

Table 3.4: The estimates for subgroup effect sizes in the successfully retrieved trees ($N = 461$) from data generated with Tree complexity = model C, $K = 80$, $\bar{n} = 80$, $\sigma_\tau^2 = 0.025$, $M = 10$, $\delta_I = 0.5$, $\mathbf{R}$ = the computed correlation matrix, Variable type = ordinal moderators.

	Fixed effect meta-CART			RE meta-CART		
	δ	averaged $\hat{\delta}$	coverage of 95% CIs	δ	averaged $\hat{\delta}$	coverage of 95% CIs
Subgroup 1	0	-0.014 (0.042)	0.835	0	-0.010 (0.053)	0.937
Subgroup 2	0	0.023 (0.058)	0.837	0	0.027 (0.059)	0.933
Subgroup 3	0.5	0.498 (0.037)	0.828	0.5	0.498 (0.044)	0.948

[6] The numbers in parentheses display the standard deviations.

3.8 Discussion

3.8.1 Conclusion, strengths, shortcomings, and remaining issues

This study proposed new strategies for the meta-CART approach of Dusseldorp et al. (2014) and Li, Dusseldorp, and Meulman (2017) as integrated procedures using the FE or RE model, and investigated the performance of the new strategies via an extensive simulation study. The simulation results show that the Type I error rates of meta-CART are mainly influenced by the number of studies included in a meta-analysis. By varying the pruning rule for different number of studies, the Type I error of meta-CART is satisfactory (≤ 0.05). The power and recovery rates of meta-CART mainly depend on the number of studies, the complexity of the true model, the type of moderator variables, the within-study sample size, the magnitude of interaction effect(s), and the residual heterogeneity. The simulation study used four tree structures with increasing complexity to assess the ability of meta-CART to retrieve the true model underlying the data set. In general, meta-CART performed well in retrieving simple models (models with a main effect or one two-way interaction effect). For more complex models (models with two two-way interaction effects or a three-way interaction effect), the power and recovery rates of meta-CART varied from low (≤ 0.10) to high (≥ 0.80) depending on the design factors.

The strength of the simulation study is that we extensively examined the influence of both observable design factors (i.e., the number of studies, the within-study sample size, the type of moderators, and the number of moderators) and unobservable design factors (i.e., complexity of the interaction effects, the magnitude of the interaction effect, the correlation between moderators, and the residual heterogeneity). These design factors covered various situations that are encountered in practice. By taking into account residual heterogeneity unexplained by the moderators, the simulation study also covers the situations that not all the influential moderators are collected in the data. The results show that the conditions resulting in high performance of meta-CART and those resulting in low performance are both encountered.

One limitation of the simulation study is that the performance of meta-CART on mixed types of moderators was not examined, although the algorithm can be applied to data sets with mixed types. The main difficulties of investigating the influence of mixed types of moderators are to define true models to generate artificial data, and to define the proportion of each type of the potential moderators in generated data sets. The large number of possible combinations of variable types and proportions within each variable type makes it difficult to extensively examine the performance of meta-CART on mixed types of moderators. In our study, the simulation results show that the Type I error rates are not largely influenced by the types of moderators, but the power and the recovery rates are strongly influenced by the type of moderators. The recovery rates are highest for binary moderators, and lowest for continuous moderators. To roughly assess the performance of meta-CART on mixed types of moderators, we performed a small simulation study by applying FE meta-CART to 1000 data sets generated with mixed types of moderators for one combination of the other design factors ($K = 80$, $\bar{n} = 80$, $\mathbf{R} = \mathbf{I}$, $M = 5$, $\sigma_\tau^2 = 0$). The true model to generate the data consists of a first split with a binary moderator, and two two-way interactions with an ordinal moderator and a nominal moderator. Given the same combination of other design factors, the estimated power rate of mixed moderators (0.994) is comparable to the estimated power rates of binary, ordinal and nominal moderators (1.000, 1.000, and 0.998, respectively). The estimated recovery rate of mixed moderators (0.812) is lower than binary moderators (0.991), but higher than ordinal and nominal moderators (0.653 and 0.649, respectively). Thus, it might be plausible to assume that the recovery rates on mixed types of moderators will be in-between the ones on binary variables and continuous variables. Future study is needed to obtain a solid conclusion about the performance of meta-CART on mixed types of moderator variables.

Another limitation is that the designed models that were used to generate data did not take linear relationship between effect size and continuous moderators into account. If the effect size is linearly related to continuous moderators, meta-CART will have difficulties to decide the split points, which is a well-known disadvantage of tree-based methods (Friedman, Hastie, & Tibshirani, 2001). One way to solve this problem is to first adjust for the linear relationship (e.g., fit a meta-regression model with main effects of continuous moderators), and then fit a meta-CART model using the adjusted effect size (i.e., the residuals from the first step) as the response variable. Furthermore, for data generated from the designed tree models, meta-CART has lower recovery rates for continuous moderators than binary, nominal, and ordinal moderators. This might be because the greedy search algorithm of meta-CART may mistakenly select a local spike when the number of possible split points to be evaluated is large. One possible solution can be using a smooth function to approximate the threshold indicator, for example, the sigmoid smooth function (see Su et al., 2016). It will be interesting to improve the performance

of meta-CART for continuous moderators for both linear and non-linear relationship in future.

A final limitation is that the simulation study did not examine the coverage of the confidence intervals of the effect size estimates for all combinations of the design factors due to the computation cost and the difficulty to compare subgroups for equivalent trees with different expressions (for an example, see in Supporting Material Figure 3.7). The analysis results from one cell of the design showed that FE meta-CART results in too narrow confidence intervals while RE meta-CART results in confidence intervals with coverage close to the nominated probability. This is because FE meta-CART ignores the uncertainty introduced by the residual heterogeneity.

One advantage of meta-CART is that it can deal with multiple moderators and identify interactions between them. In addition, the simulation results show that the performance of meta-CART is not (largely) influenced by the number of moderators, and the correlation between the moderators. Meta-CART also has the potential to be extended and integrated into other advanced meta-analytic techniques such like multiple group modeling (Schoemann, 2016) and meta-analytic structural equation modeling (Cheung & Chan, 2005). Multiple group modeling is a powerful tool for testing moderators in meta-analysis, but it can only be used to test for categorical moderators; continuous moderators cannot be assessed with this technique. Meta-CART can create a subgrouping variable based on continuous moderators, which could be used as a categorical moderator to be tested in a multiple group model. Meta-analytic structural equation modeling (MASEM) is an increasingly popular technique for synthesizing multivariate correlational research. An extended approach of MASEM by Wilson, Polanin, and Lipsey (2016) uses meta-regression to generate covariate-adjusted correlation coefficients for input to the synthesized correlation matrix capable of reducing the influence of selected sources of heterogeneity. Since meta-CART can be used to identify multiple moderators that account for the sources of heterogeneity, it will be interesting to incorporate meta-CART into MASEM and evaluate the performance in future work. A final advantage is that meta-CART can keep good control of Type I error (≤ 0.05) by the pruning procedure with cross-validation. Higgins and Thompson (2004) observed high rates of false-positive findings from meta-regression as it is typically practiced. They found that the Type I error rate of FE meta-regression is unacceptable in the presence of heterogeneity. In addition, the Type I error problems are compounded for both FE and RE meta-regression when multiple moderators are assessed. Compared to meta-regression, FE meta-CART has acceptable Type I error rates even in presence of residual heterogeneity. And the Type I error rates are not largely influenced by the number of moderators for both FE meta-CART and RE meta-CART.

Although meta-CART has a good control of the risk of spurious findings, caution is needed for the interpretation of the significance test. Because the moderator effects are explored and tested on the same data set, the test based on the between-subgroups

Q-statistic is a pseudo Q-test. The test only gives information when the moderator effects are not significant. But it does not confirm the significance of the moderator effects. The Type I error of meta-CART is mainly controlled by the pruning procedure rather than the Q-test. Therefore, meta-CART should be used as a hypothesis generating tool (i.e., an exploratory method) rather than a hypothesis testing method. To confirm the generated hypothesis, standard meta-analytical methods such as meta-regression and subgroup meta-analysis can be employed to test the influential moderator (interaction) effects on new data.

An interesting phenomenon is that FE meta-CART had higher recovery rates for model E than model D, but RE meta-CART showed the opposite. A possible explanation is the difference between model D and the models A, B, C, and E. In contrast to other tree models, the tree size of model D is sensitive to the first splitting variable that the algorithm chooses. For example, if the algorithm chooses the moderator x_2 instead of x_1 as the first splitting variable, the final tree will end up with six instead of four terminal nodes. An illustration of the two equivalent trees can be found in Supporting Material Figure 3.7. Since RE meta-CART employs the sequential partitioning procedure, it could be more sensitive to the order of the chosen splitting variables than FE meta-CART, which employs a recursive partitioning procedure.

As a recursive partitioning method and a sequential partitioning method respectively, both FE meta-CART and RE meta-CART use local optimization procedures. Thus, the algorithm may find a local optimum solution rather than a global optimum. For example, when applying to the illustrative data set (see section 3.2), FE meta-CART results in a local optimum solution with "T1: provide information about behavior-health link" being the first splitting variable and "T4: prompt intention formation" being the second splitting variable. However, if we apply a "look-ahead" procedure that searches through all possible combinations of two splitting variables on the same data set, the resulting solution will have "T4: prompt intention formation" as the first splitting variable and "T1: provide information about behavior-health link" as the second splitting variable. It results in a higher FE between-subgroups Q-statistic ($Q_B = 41.78$) compared to the FE meta-CART solution ($Q_B = 40.59$). To overcome this local optimum problem, one promising improvement of meta-CART would be to develop a global optimization algorithm. Such an algorithm for both FE and RE models can improve meta-CART from several aspects: 1) to avoid local optimum solutions, 2) to reduce the sensitivity of RE meta-CART to the sequence of the partitioning as mentioned above, 3) to make the two different partitioning procedures of FE meta-CART and RE meta-CART more similar. Thus, it will be worthwhile to develop a global optimization method in future work.

In this study, the ordinal and nominal moderators were generated with three levels. Because in meta-analytic practice most ordinal moderators commonly have three levels such like "low", "medium" and "high", and categorical moderators commonly have two

or three levels, we did not examine the performance of meta-CART on moderators with larger number of levels. If there are moderators with different numbers of levels, the greedy search property of meta-CART might induce a selection bias towards variables that have more possible split points (Doyle, 1973). A solution to address this selection bias is to adapt the GUIDE (Generalized, Unbiased Interaction Detection and Estimation) algorithm by Loh (2002) to the framework of meta-CART.

3.8.2 The guideline for application of meta-CART

Based on the simulation results, the recommended pruning rule (expressed as a c*SE rule) depends on the type of research at hand and the number of studies. A higher value of c indicates more pruning. If higher power and recovery rates are more important than strict control of Type I error for a specific research problem, a smaller pruning parameter c can be used. For example, researchers may apply meta-CART with a liberal pruning rule using $c = 0$ or 0.5 to gain more power by risking higher Type I error rates. If a strict control of the Type I error (≤ 0.05) is required, a stricter pruning rule using $c = 1$ can be applied when the number of studies $K < 80$, and $c = 0.5$ when $K \geq 80$ for FE meta-CART. For RE meta-CART, the pruning rule $c = 1$ can be used when $K < 120$ and $c = 0.5$ when $K \geq 120$. To perform a meta-CART analysis with satisfactory performance (i.e., with power and recovery rates both higher than 0.80), a minimum number of studies $K = 40$ is required to detect main effect or simple interaction effect such as one two-way interaction, and $K = 80$ is required to detect more complex interaction effects.

The choice of whether to use FE or RE meta-CART should be based on the assumption of the residual heterogeneity and the research question, but not on the power and the recovery rates. General discussion and guidelines about the choice between FE model and RE model in meta-analysis can be found in Borenstein, Hedges, Higgins, and Rothstein (2010), and Schmidt, Oh, and Hayes (2009). For meta-CART analysis, heterogeneity is likely to exist when the number of studies is large (i.e., $K \geq 40$). In addition, FE meta-CART may result in over-optimistic confidence intervals when residual heterogeneity exists. To conclude, RE meta-CART is generally recommended, unless there is a priori grounding for the fixed effect assumption.

3.9 Supporting Materials

3.9.1 The effect sizes of the design factors

The partial η^2s of the design factors in the ANOVA are shown in Tables 3.5-3.10. Only the ten most influential factors are shown for each analysis.

Table 3.5: Five most influential factors in the ANOVA on Type I error rate for FE meta-CART

The name of predictor variable	generalized partial η^2
K	0.70
c	0.39
$K \times c$	0.06
M	0.04
σ^2_τ	0.03

Table 3.6: Five most influential factors in the ANOVA on Type I error rate for RE meta-CART

The name of predictor variable	generalized partial η^2
$K \times c$	0.57
K	0.27
Variable type	0.04
c	0.04
M	0.03

3.9.2 Two representations for model D

In the simulation study, a model with two two-way interactions (model D) was used to generate data with complex interaction effects. This tree model can be represented in two trees with different number of splits (see Figure 3.7). The number of the splits depends on the first splitting variable.

Table 3.7: Ten most influential factors in the ANOVA on power rate for FE meta-CART

The name of predictor variable	partial η^2
Tree complexity	**0.97**
δ_I	**0.95**
K	**0.92**
σ_τ^2	**0.85**
$\bar{n}$	0.75
Variable type	0.71
Variable type × Tree complexity	0.60
$K \times \delta_I \times$ Tree complexity	0.52
Variable Types $\times \delta_I \times$ Tree complexity	0.50
$\delta_I \times \sigma_\tau^2$	0.49

Table 3.8: Ten most influential factors in the ANOVA on power rate for RE meta-CART

The name of predictor variable	partial η^2
Tree complexity	**0.98**
δ_I	**0.96**
K	**0.91**
Variable type	**0.88**
σ_τ^2	**0.86**
$\bar{n}$	**0.82**
$K \times \sigma_\tau^2 \times$ Tree complexity	0.64
Variable type × Tree complexity	0.64
$\delta_I \times \sigma_\tau^2$	0.44
$\delta_I \times$ Tree complexity	0.43

Table 3.9: Ten most influential factors in the ANOVA on recovery rate of moderators for FE meta-CART

The name of predictor variable	partial η^2
Tree complexity	**0.99**
K	**0.96**
δ_I	**0.95**
σ_τ^2	**0.88**
Variable type	**0.85**
$\bar{n}$	0.79
M	0.70
Variable type × Tree complexity	0.70
$\delta_I \times \sigma_\tau^2$	0.63
$K \times$ Tree complexity	0.60

Table 3.10: Ten most influential factors in the ANOVA on recovery rate of moderators for RE meta-CART

The name of predictor variable	partial η^2
Tree complexity	**0.99**
δ_I	**0.96**
K	**0.94**
Variable type	**0.90**
σ^2_τ	**0.89**
$\bar{n}$	**0.85**
Variable type× Tree complexity	0.73
M	0.63
$\delta_I \times \sigma^2_\tau$	0.55
δ_I× Tree complexity	0.50

Table 3.11: Ten most influential factors in the ANOVA on difference in average recovery rates between FE meta-CART and FE meta-regression

The name of predictor variable	partial η^2
Tree complexity	**0.96**
K	0.69
σ^2_τ	0.65
Variable type	0.62
$K \times \delta_I$	0.58
M	0.48
$K \times \bar{n}$	0.35
$\bar{n}$	0.19
δ_I	0.18
R	0.01

Table 3.12: Ten most influential factors in the ANOVA on difference in average recovery rates between RE meta-CART and RE meta-regression

The name of predictor variable	partial η^2
Tree complexity	**0.93**
Variable type	0.76
M	0.31
$K \times \delta_I$	0.30
$K \times \bar{n}$	0.09
K	0.08
δ_I	0.03
$\bar{n}$	0.01
R	0.00
σ^2_τ	0.00

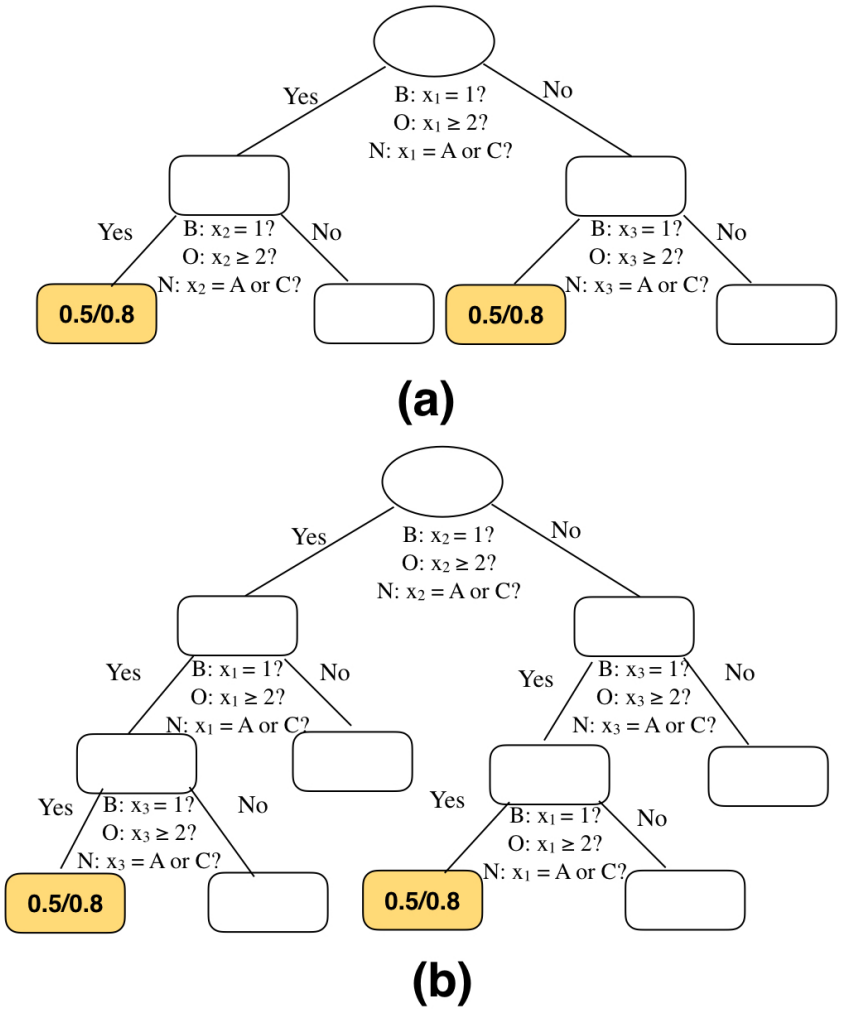


Figure 3.7: Two equivalent expressions for model D. The different number of splits depend on the first splitting variable.

Chapter 4

Multivariate moderator meta-analysis with the R-package metacart

abstract

In meta-analysis, heterogeneity often exists between studies. Knowledge about study features (i.e., moderators) that can explain the heterogeneity in effect sizes can be useful for researchers to assess the effectiveness of existing interventions and design new potentially effective interventions. When there are multiple moderators, they may amplify or attenuate each other's effect on treatment effectiveness. However, in most meta-analysis studies, interaction effects are neglected due to the lack of appropriate methods. The method meta-CART was recently proposed to identify interactions between multiple moderators. The analysis result is a tree model in which the studies are partitioned into more homogeneous subgroups by combinations of moderators. This paper describes the R-package `metacart`, which provides user-friendly functions to conduct meta-CART analyses in R. This package can fit both fixed- and random-effects meta-CART, and can handle dichotomous, categorical, ordinal and continuous moderators. In addition, a new look ahead procedure is presented. The application of the package is illustrated step-by-step using diverse examples.

4.1 Introduction

Methodology for synthesizing findings from multiple studies addressing the same research question has a long history (Hedges, 1981; Hedges & Olkin, 1985). The typical goals of meta-analysis are to estimate the overall effect size (i.e., a weighted average of study effect sizes), to quantity the heterogeneity in the study effect sizes, and to investigate the study characteristics that explain the heterogeneity (i.e., moderators). The relationship between moderators and the study effect sizes can be of high interest for behavioral scientists to evaluate existing interventions and to design new potentially effective interventions. In the meta-analysis framework, moderator analysis can be conducted with several stand-alone software programs such as **Comprehensive Meta-Analysis** (Borenstein et al., 2009), **Meta-Analyst** (Wallace, Schmid, Lau, & Trikalinos, 2009), and the Cochrane Collaboration's **RevMan** (Collaboration, 2008). Also, there are add-ins/macros/packages that can be used to conduct moderator analysis in many software languages such as **Stata** (StataCorp 2017, for details, see Sterne, Bradburn, and Egger 2008), **SAS** (Inc 2002, for details, see Wang and Bushman 1999, Arthur Jr, Bennett, and Huffcutt 2001, and Sheu and Suzuki 2001), **SPSS** (Corp 2013, for details, see Lipsey and Wilson 2001), and **R** (Team 2017, e.g., package **rmeta** by Lumley 2012, **mvmeta** Gasparrini, Armstrong, and Kenward 2012, and **metafor** by Viechtbauer 2010). Most of these programs/add-ins are based on meta-regression, which is the most commonly used method for moderator analysis in meta-analysis.

In practice, behavioral scientific meta-analyses often have multiple moderators to be examined. For example, interventions to change health-related behavior generally include various behavior change techniques (BCTs), and researchers are interested in investigating the influence of BCTs on the effectiveness of interventions (see Michie, Abraham, et al., 2009). However, meta-regression has several limitations in such cases. First, when the number of included studies is small, meta-regression suffers from low statistical power to examine all moderators simultaneously (Tanner-Smith & Grant, 2018). Second, meta-regression has difficulties in exploring interaction effects among moderators since it requires moderators and their interactions to be specified beforehand. When there is no a priori hypotheses available, the number of all possible interaction terms are usually too large to be included in one model. The interaction effects, however, can provide valuable information to answer questions such like "does these intervention components amplify or attenuate each other's effectiveness?" and "which combination of study characteristics results in the highest effectiveness?".

To overcome the aforementioned limitations, tree-based models can be integrated into the framework of meta-analysis. Tree-based methods were introduced for the first time by Morgan and Sonquist (1963) in a method called automatic interaction detection (AID), and were fully developed in classification and regression trees (CART) by Breiman et al.

(1984). Trees are good at dealing with many predictor variables that may interact, and produce results that can be easily interpreted. Tree-based methods have been used in the field of behavioral and medical sciences (e.g, Finch et al., 2011; Leach et al., 2016; Trujillano, Badia, Serviá, March, & Rodriguez-Pozo, 2009), but the idea of using trees in the meta-analysis framework is relatively new. For individual patient data (IPD) meta-analyses, Mistry, Stallard, and Underwood (2018) proposed a recursive partitioning method called IPD-SIDES that identifies patient subgroups by individual characteristics that may be related to the response to intervention. For aggregated data meta-analyses, a method called meta-CART was proposed to identify interaction effects among study-level characteristics (Dusseldorp et al., 2014; Li, Dusseldorp, & Meulman, 2017, 2019). This paper focuses on moderator analyses on aggregated meta-analysis by meta-CART. Compared to meta-regression, meta-CART has several advantages: first, it excels at dealing with interaction effects, and the interactions can be easily interpreted; second, it has automatic variable selection and does not require model selection; third, it is able to handle non-linear associations between moderators and effect size. Furthermore, since tree models are invariant to monotone transformation of predictors, meta-CART can keep the ordering information of ordinal moderators, whereas meta-regression usually codes ordinal variables the same as categorical variables. Li et al. (2019) showed via a simulation study that meta-CART can achieve satisfactory power and recovery rates (i.e., ≥ 0.80) with a sufficiently large sample size (i.e., $n \geq 40$ for simple interactions and $n \geq 80$ for complex interactions). The meta-CART method has been acknowledged as a potential alternative statistical method for meta-regression to understand the combined effects of moderators (Michie et al., 2015; O'Brien et al., 2015; Tipton et al., 2018), and it has been applied in several meta-analytic studies (e.g., Bull et al., 2018; van Genugten, Dusseldorp, Webb, & van Empelen, 2016).

In the present paper, we introduce the R package **metacart**, which implements the meta-CART method. This package provides user-friendly functions to perform meta-CART analysis for various types of moderators (i.e., continuous, ordinal, and categorical variables), and includes various additional options such as tuning the pruning of the tree model, restricting the minimum number of studies in a subgroup, and so on. In addition, we developed a new option to apply a "look-ahead" strategy specifically focusing on interaction detection. This paper aims to provide a general overview of the capabilities of the **metacart** package for conducting multivariate moderator meta-analysis with R. Section 2 introduces the meta-CART method. Section 3 describes the main functions in the **metacart** package, and Section 4 illustrates their practical usage with examples of real meta-analyses. Section 5 contains concluding remarks.

4.2 Meta-CART Method

4.2.1 The goal of meta-CART analysis

The underlying goal of meta-CART analysis is to identify subgroups defined by the moderators that can explain the heterogeneity in the study effect sizes. Appropriate data for meta-CART include an outcome variable of interest (i.e., study effect size), the within-study sample variance of the effect size, and the potential moderator variables that may influence the effect size. Depending on the type of study, there is a variety of different effect size measures, including the odds ratio, the relative risk, the correlation coefficient, and (the standardized) mean difference. Meta-CART can deal with all these effect size measures as long as all the studies are using the same measure[1]. To identify influential moderators, meta-CART partitions the studies into subgroups that are more homogeneous with respect to their study effect sizes. The result is a tree model with the terminal nodes as the identified subgroups and the splitting variables as the influential moderators. Figure 4.1 shows an example of a tree model fitted by the package **metacart**. The root node (i.e., the node at the top) represents all the studies that are included in the analysis. From the root node, the studies are partitioned into two subgroups by applying a threshold on the values of a moderator. For example, the root node in Figure 4.1 is partitioned into two child nodes based on the moderator "eye_assess". This moderator is categorical, and the studies of which "eye_assess" equals to "E.1.a.ii", "E.1.b", or "E.1.c" fall into the left child node and the other studies fall into the right child node. If the moderator is an ordinal or a continuous variable, a binary question such as "is the value of the moderator smaller than the split point?" will be asked to introduce a split. Note that the moderators and the corresponding split points are automatically selected by the algorithm. How the moderators and split points are selected will be explained in Section 4.2.2.

The resulting subgroup memberships are defined by a hierarchically nested set of decision rules such as "does the intervention include certain treatment components?" or "is the average age of the patients smaller than 25?". In the identified subgroups, the summary effect sizes and their confidence intervals are estimated. The analysis results can be used by researchers to predict the effect size given the study characteristics, or to identify the combination of study characteristics that results in the highest/lowest effect size. Note that meta-CART has an exploratory nature, since the subgroups are not predefined but identified from the data. Thus, meta-CART should be used as a hypothesis-generating tool, and we recommend testing the identified moderators by confirmatory methodology in further studies.

[1]Most effect size measures can be converted to one another (see Hedges & Olkin, 1985).

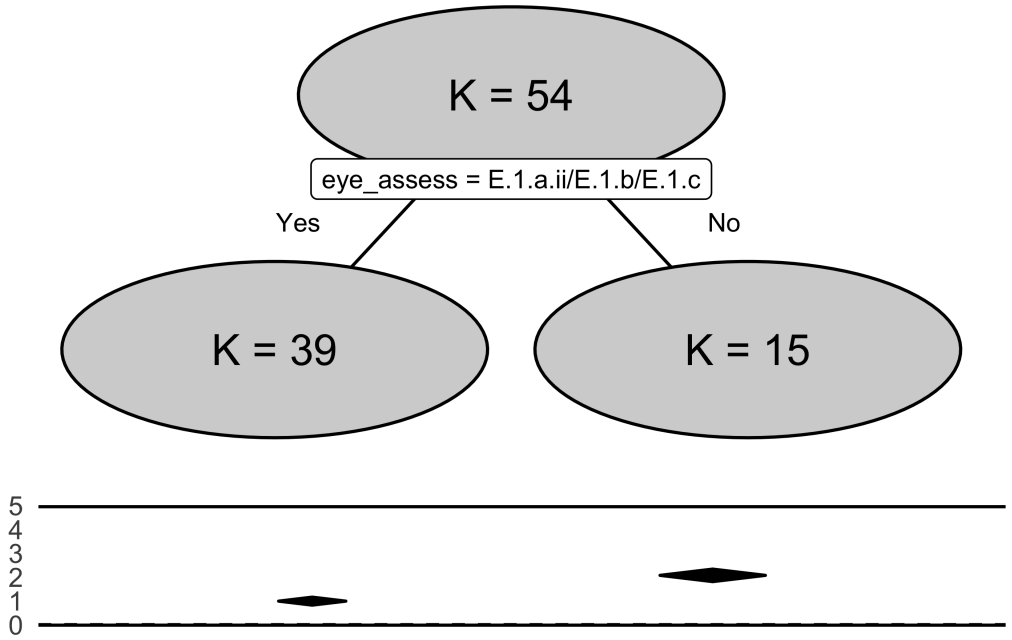


Figure 4.1: The meta-CART analysis results of 54 samples examining the influence of moderators on the association between handedness and eye-dominance. The figure shows the main effect of the method to assess eye-dominance, which partitions the samples in two subgroups. The two solid lines show the range of the effect sizes of all the studies. The diamonds between the solid lines present the 95% confidence intervals of the summary effect sizes.

4.2.2 Meta-CART algorithm

There are two types of meta-CART algorithms: fixed-effect (FE) meta-CART, a recursive partitioning algorithm that ignores the residual heterogeneity unexplained by the moderators, and random-effects (RE) meta-CART, a sequential partitioning algorithm that takes into account the residual heterogeneity. As in standard meta-analysis, the choice between FE and RE assumptions should be based on a researcher's prior belief; in other words, a researcher should decide on which assumption to use before analyzing the data. In general, if a researcher believes that given the study characteristics the variance of the study effect sizes is merely due to the sampling variance[2], FE meta-CART can be chosen. If there is no a priori information about the residual heterogeneity, RE meta-CART is recommended. A more general discussion about FE model and RE model in meta-analysis can be found in Borenstein et al. (2010) and Schmidt et al. (2009).

In this section, we describe the FE and RE meta-CART algorithms, which underly the **metacart** package.

[2]This implies that differences between the studies are all accounted for by moderators, and sampling variance accounts for the variance of the study effect sizes adjusted for moderator effects.

Fixed effect meta-CART

FE meta-CART splits the studies into more homogeneous subgroups under the FE assumption; that is, moderator effects and the within-study sampling variance are the only two sources of the variation in study effect sizes. Denote the true effect size in the k^{th} study by δ_k, and denote the observed effect size in the k^{th} study by d_k.[3] Under the FE assumption, the observed effect size is given by

$$d_k = \delta_k + \epsilon_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_M x_{Mk} + \epsilon_k, \tag{4.1}$$

where x_{mk} $(m = 1, ..., M)$ specify the values on the M moderators of the k^{th} study, and the βs are the corresponding coefficients. The sampling error ϵ_k is assumed to be distributed as $\mathcal{N}(0, \sigma^2_{\epsilon_k})$, where $\sigma^2_{\epsilon_k}$ is the sampling variance.

The summary effect size is computed as the weighted mean, with weights $w_k = 1/\sigma^2_{\epsilon_k}$:

$$d_+ = \frac{\sum d_k/\sigma^2_{\epsilon_k}}{\sum 1/\sigma^2_{\epsilon_k}}. \tag{4.2}$$

The measure of heterogeneity, the Q-statistic, is given by

$$Q = \sum_{k=1}^{K} \frac{(d_k - d_+)^2}{\sigma^2_{\epsilon_k}}. \tag{4.3}$$

Starting from one group including all the studies (i.e., the root node), FE meta-CART partitions the root node into two subgroups (i.e., offspring nodes), by searching through all possible splits and finds the moderator with corresponding split point that maximizes the between-subgroups heterogeneity. Denote the summary effect size of the t^{th} subgroup (i.e., t^{th} node in the tree) by d_{t+}, the between-subgroups heterogeneity measure is computed as

$$Q_B = \sum_{t}^{|T|} \sum_{k \in t} \frac{(d_{t+} - d_{++})^2}{\sigma^2_{\epsilon_k}}, \tag{4.4}$$

where $|T|$ is the total number of subgroups[4], and d_{++} is the weighted grand mean of the parent node.

To grow a tree, FE meta-CART recursively searches the split that maximizes the heterogeneity Q_B between the left and right child nodes. After each split, the algorithm partitions a parent node into two child nodes. The tree-growing process is a recursive partitioning procedure since the same operation can be applied to any child node itself

[3]Note that we focus here on the d-family of effect sizes, that is, measures of the standardized mean outcome difference between treatment and control groups (e.g., Cohen's d or Hedges' g). However, meta-CART can also be used for other effect size measures.

[4]$|T| = 2$ when only considering the heterogeneity between left and right child nodes after a split of the parent node.

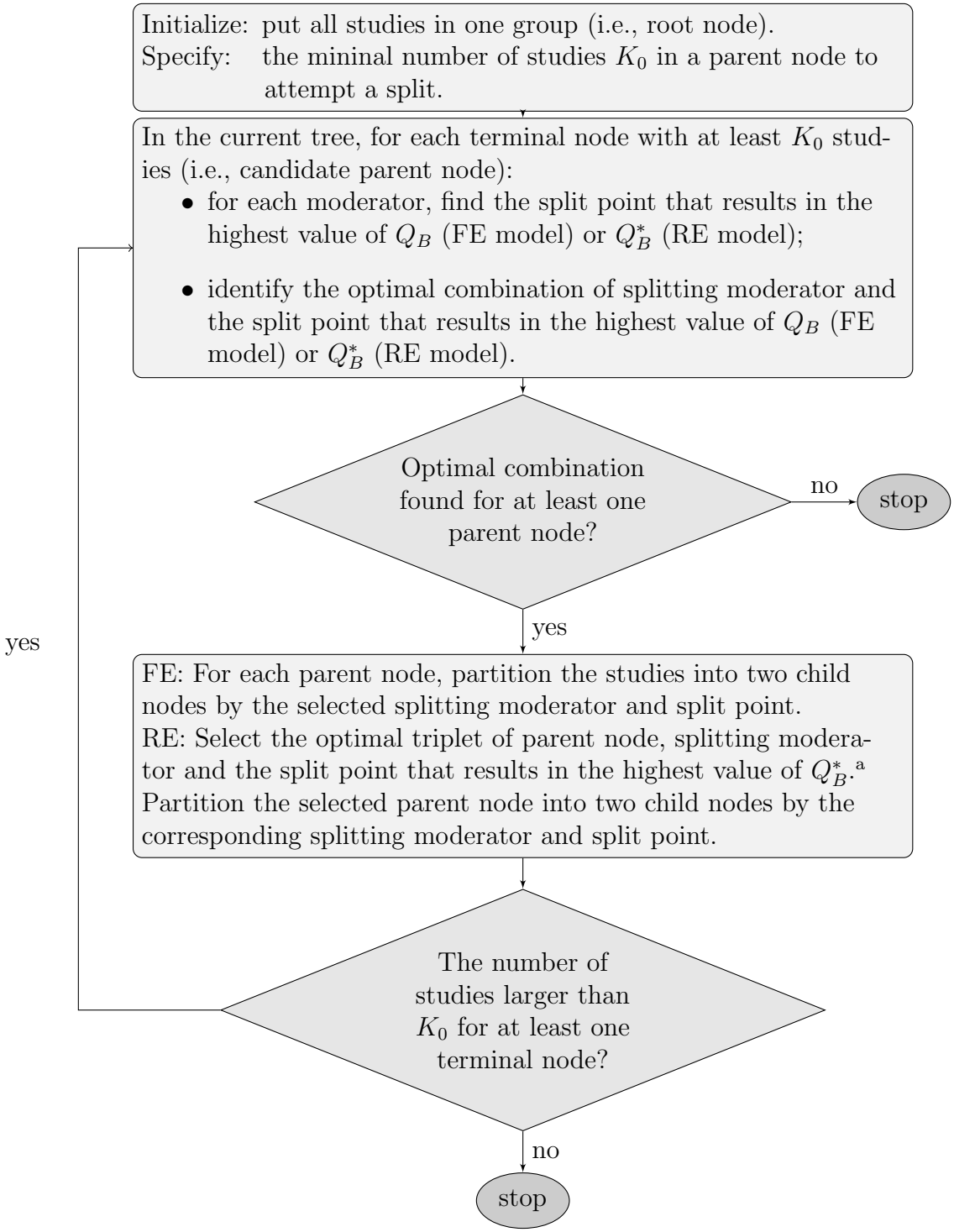


Figure 4.2: Flowchart of tree growing process of meta-CART algorithms for fixed-effect (FE) model and random-effects (RE) model.

[a] Note that this substep is not needed for recursive partitioning method like ordinary CART and FE meta-CART, but necessary for RE meta-CART due to the re-estimation of σ_τ^2.

without taking the other nodes in the current tree into account. A flowchart of the tree growing process of FE meta-CART is shown in Figure 4.2. The splitting process continues until all terminal nodes contain fewer than K_0 studies, where K_0 is a user-specified threshold (see section 4.3). To prevent overfitting, the initial tree will be iteratively pruned to a nested sequence of subtrees, from which a best-sized tree is chosen via cross-validation. Different rules can be applied to determine the best-sized tree. To generalize the rules, a parameter c is introduced to select the "optimal" tree by using the $c \cdot SE$ rule (Dusseldorp et al., 2010). The $c \cdot SE$ rule selects the smallest tree with cross-validation error within the minimum cross-validation error plus its standard error multiplied by c.

The final analysis results consist of the selected tree model that represents the interactions between moderators, the estimates of the summary effect sizes d_{t+} within each subgroup, and the between-subgroups Q_B with $df = |T| - 1$. It should be noted that the significance test based on the between- subgroups Q-statistic is an overoptimistic pseudo Q-test. Given pre-defined subgroup membership, the between-subgroups Q-statistic follows a chi-square distribution with $df = |T| - 1$ under the null hypothesis that there is no significant heterogeneity between the subgroups. However, in meta-CART the subgroup membership is not pre-defined, but identified by the tree-growing and cross-validation procedures. Therefore, over-optimism exists in the Q-test, and the p-value should not be interpreted as the face value. It only gives information when the moderator effects are not significant. In other words, a non-significant p-value indicates that the influence of the identified moderator(s) on study effect sizes is not significant, but a significant p-value does not confirm the significance of the moderator effects. It is also worth to note that despite the over-optimism in the Q-test, the Type I error rate of meta-CART (i.e., defined as the rate of finding a nontrivial tree with a significant between-subgroups Q-statistic while there is no moderator effect in the true structure underlying the data) is not inflated, because the Type I error of moderator effects is mainly controlled by the pruning procedure (for details, see Li et al., 2019).

RE meta-CART

RE meta-CART takes the residual heterogeneity unexplained by moderators into account. The RE model is expressed as below:

$$d_k = \beta_0 + \beta_1 x_{1k} + \beta_2 x_{2k} + ... + \beta_M x_{Mk} + \tau_k + \epsilon_k, \tag{4.5}$$

where τ_k is the variation introduced by the residual heterogeneity. Denote the residual heterogeneity by σ_τ^2, with τ_k distributed as $\mathcal{N}(0, \sigma_\tau^2)$. There are various estimators for σ_τ^2, including the Hunter-Schmidt estimator (Schmidt & Hunter, 2014), the Hedges estimator (Hedges & Olkin, 1985), the DerSimonian-Laird estimator (DerSimonian & Laird, 1986), the Sidik-Jonkman estimator (Sidik & Jonkman, 2005), and the maximum-likelihood or

restricted maximum-likelihood estimator (Viechtbauer, 2005). In the **metacart** package, we use the DerSimonian-Laird estimator for its lower computational cost.[5]

With the estimated residual heterogeneity, the summary effect size can be computed with the RE weights $w_k^* = 1/(\sigma_{\epsilon_k}^2 + \sigma_\tau^2)$:

$$d_+^* = \frac{\sum d_k/(\sigma_{\epsilon_k}^2 + \sigma_\tau^2)}{\sum 1/(\sigma_{\epsilon_k}^2 + \sigma_\tau^2)}. \tag{4.6}$$

The RE heterogeneity is given by

$$Q^* = \sum_{k=1}^{K} \frac{(d_k - d_+^*)^2}{\sigma_{\epsilon_k}^2 + \sigma_\tau^2}. \tag{4.7}$$

Similar to FE meta-CART, RE meta-CART starts from the root node, and searches for the split that maximizes the between-subgroups heterogeneity. The difference is that before searching for the split, RE meta-CART needs to estimate the residual heterogeneity (σ_τ^2) first, since the RE between-subgroups heterogeneity is given by

$$Q_B^* = \sum_{t}^{|T|} \sum_{k \in t} \frac{(d_{t+}^* - d_{++}^*)^2}{\sigma_{\epsilon_k}^2 + \sigma_\tau^2}. \tag{4.8}$$

To continue the splitting process, RE meta-CART updates the estimate for σ_τ^2 and searches for the new split maximizing the partitioning criterion. A flowchart of the tree-growing process of RE meta-CART is also given in Figure 4.2. Note that a split of a node will globally affect the estimation of σ_τ^2 and the value of Q_B^*. Thus, in contrast to FE meta-CART, RE meta-CART considers heterogeneity between all the terminal nodes rather than only between the resulting left and right child nodes after a split. As a result, the tree-growing process of RE meta-CART is not fully recursive, since the algorithm needs to take all terminal nodes of the current tree into account to introduce a new split. In other words, the estimate of σ_τ^2 and the optimal choice for a new split depend on the sequence of previous splits. Instead, RE meta-CART applies a sequential partitioning algorithm. [6]

As in FE meta-CART, the splitting process of RE meta-CART continues until a large tree is grown. Then an optimally-sized subtree is selected using cross-validation with the $c \cdot SE$ rule. The associated between-subgroups Q_B^*, the estimates for residual heterogeneity σ_τ^2, and the within-subgroup summary effect sizes d_{j+}^* are obtained as the final tree is selected.

[5] In our pilot simulation study, it was found that the different choices of the estimator led to similar trees.

[6] A sequential tree-growing algorithm is defined as an algorithm that only splits a node in the tree if the reduction in the partitioning criterion at the proposed node exceeds some value, regardless of what might happen lower in the tree (**?**, Chapter 4). Recursive partitioning can be seen as a restricted form of sequential partitioning where at each node, exactly the same splitting procedure is repeated.

Look-ahead strategy for RE meta-CART

The growing algorithms of both FE and RE meta-CART are fully greedy. Although the approach guarantees a locally optimal solution at each split, it does not guarantee a globally optimal solution for the whole tree. That is, the tree-growing procedure chooses each optimal split with no regard of future splits. Because the estimated residual heterogeneity is influenced by the sequence of partitioning, the RE meta-CART is more sensitive to this local optimization problem. To alleviate this problem, we propose a look-ahead strategy for RE meta-CART, which examines two steps ahead instead of one at the split of the root node. This look-ahead strategy is applied only to the top level of a tree.

Starting the algorithm at the root node, a standard RE meta-tree chooses the single split (i.e., one split point on one moderator) that maximizes the between-subgroups heterogeneity. In contrast, a look-ahead strategy searches for the combination of two splits (i.e., single split points on two moderators or two split points on one moderator) that maximizes the partitioning criterion. As a result, the entire growing procedure that applies a look-ahead strategy consists of two sub-procedures. The first sub-procedure grows a tree with two splits, which searches for the optimal combination of a split of the root node and a split of one of its child nodes. The combination that maximizes Q_B^* is chosen. In the second sub-procedure, the resulting offspring nodes of the first two splits are split following a fully greedy procedure for maximizing Q_B^* at each split. This complete splitting procedure grows a large tree, which will be pruned using the pruning procedure mentioned above.

It should be noted that such a strategy does not guarantee a globally optimal solution, and only partially addresses the local optimization issue. Although more steps to look ahead can be beneficial, we chose the number of steps as two because the computational burden of a look-ahead procedure grows exponentially as the number of steps increases (Esmeir & Markovitch, 2007).

4.3 The metacart package

The **metacart** package provides functions to perform both FE and RE meta-CART analysis. The package is available via the Comprehensive R Archive Network (CRAN) at `https://cran.r-project.org/package=metacart`, and can be directly installed within R by typing `install.packages("metacart")`. The current version is 2.0-0.

4.3.1 Functions for FE meta-CART

The function `FEmrt` is the main function to perform a fixed effect meta-CART analysis. Both the tree growing and pruning processes in the function `FEmrt()` are based on the function `rpart()` in the **rpart** package (Therneau, Atkinson, & Ripley, 2017), with adap-

tations to fit a tree model applying appropriate weights on meta-analytic data (see section 4.2.2) and automatically compute subgroup meta-analysis results within the function. We now describe the different arguments of the function.

```
FEmrt(formula, data, vi, subset, c = 1, control = rpart.control(
minsplit = 6, minbucket = 3, cp = 1e-04, xval = 10), ...)
```

The argument `formula` defines the outcome variable (i.e., effect size) and the predictor variables (i.e., moderators).

The argument `data` specifies the name of the data set to be analyzed. The argument `vi` requires the column name of the sampling variance of the effect size.

The argument `subset` is an optional expression to select a subset of the data to fit the model. The subset argument can either be a logical or a numeric vector indicating the indices of rows (i.e. studies) to be included.

The argument `c` refers to the pruning parameter (described in section 4.2.2) to be used for the analysis. The default value $c = 1$ corresponds to the one-standard-error rule recommended by Breiman et al. (1984). For meta-CART analysis, the recommended value of c depends on the type of research at hand and the number of studies. In general, if a strict control of the Type I error (less than or equal to 0.05) is required, a pruning rule using $c = 1$ can be applied when the number of studies $K < 80$, and $c = 0.5$ when $K \geq 80$ (for more guidelines about the value of the pruning parameter c, see Li et al. (2019))

The argument `control` specifies the options to control the growing and pruning processes. As mentioned above, both the tree growing and pruning processes are based on the function `rpart`, and therefore the control argument should be an `rpart.control` object (see details in Therneau et al. 2017). Within the `rpart.control` object, `minsplit` specifies the minimum number of studies that must exist in a parent node for a split to be attempted; `minbucket` specifies the minimum number of studies in any terminal node; `cp` specifies the minimal improve of complexity parameter (i.e., Q_B divided by Q) to make a split in the growing process; `xval` specifies the number of cross-validations in the pruning process. The default value of `minbucket` is chosen as 3, because a sample size of one or two is too small to produce reliable subgroup meta-analysis results. Consequently, the default value of `minsplit` is chosen as 6. The default value `xval` = 10 corresponds to the ten-fold cross-validation recommended by Breiman et al. (1984).

The output of the `FEmrt` function is an S3 object of class `"FEmrt"`. The corresponding print and summary methods can be used to display and inspect the elements of the

object. The plot method can be used to present the main effects of identified moderators and interaction effects between them in a tree model. The predict method of the S3 object allows the user to predict effect size given the value of moderators. Examples to apply these methods will be described in section 4.4.

4.3.2 Functions for RE meta-CART

The function `REmrt` is the main function to perform random effects meta-CART analysis. Both the tree growing and pruning processes in the function `REmrt` are entirely new **R**-code combined with **C++** that implements the sequential partitioning algorithm described in section 4.2.2. The different arguments of the function `REmrt` are described below.

```
REmrt(formula, data, vi, c = 1, maxL = 5, minsplit = 6, minbucket = 3,
cp = 1e-4, xval = 10, lookahead = FALSE, ...)
```

The arguments `formula`, `data`, `vi`, and `c` are similar to the corresponding arguments in the function `FEmrt`.

The argument `maxL` is an option to define the maximum number of splits in the tree growing process.

The arguments `minsplit`, `minbucket`, `cp`, and `xval` are other options to control the growing and pruning processes. These options are not given in an `rpart.control` object, because the `REmrt` is an entirely new function and does not depend on the `rpart` function. The specifications of these options are the same as those from the function `FEmrt`.

The argument `lookahead` is a logical indicator to specify whether to apply the look-ahead strategy described in section 4.2.2.

The output of the `REmrt` function is an S3 object of class `"REmrt"`. Similar to `FEmrt`, there are corresponding print, summary, plot, and predict methods for `"REmrt"` objects . Examples will be described in section 4.4.

4.4 Examples

4.4.1 New approach meta-CART compared to meta-regression

In this example, we re-analyze the data `"dat.bourassa1996"` included in the **metafor** package (Viechtbauer, 2010). The data set contains the meta-analytic data from Bourassa, McManus, and Bryden (1996), including results from 47 studies on the association between handedness and eye-dominance. Some studies included multiple (independent) samples,

resulting in 54 samples in total. Furthermore, for some studies, the combined data of the males and females were further broken down into the two subgroups. As a result, the data set contains 96 (sub)samples in total. We only selected the independent samples with the combined data of both males and females ($K = 54$). The results of each of these samples were given in terms of the number of left-handed left-eyed, left-handed right-eyed, right-handed left-eyed, and right-handed right-eyed individuals. We use the log odds-ratio as the measure of effect size, which is the same as in Bourassa et al. (1996). A higher log odds-ratio indicates a higher association between handedness and eye dominance. First, we compute the effect size and sampling variance by using the `escalc` function in the **metafor** package.

```
R> library(metafor)
R> data("dat.bourassa1996")
R> dat.handedness <- escalc(measure = "OR", subset = (sex == "combined"),
+                           ai = lh.le, bi = lh.re, ci = rh.le, di = rh.re,
+                           data = dat.bourassa1996, add = 1/2, to = "all")
R> head(dat.handedness)
```

```
  study sample author year selection investigator   hand_assess
1     1      1  Mills 1925        no        other questionnaire
2     2      2 Downey 1927       yes psychologist questionnaire
5     3      3  Miles 1930        no psychologist questionnaire
6     4      4 Quinan 1931        no        other   performance
7     5      5 Jasper 1932       yes psychologist questionnaire
8     6      6   Lund 1932        no psychologist   performance
  eye_assess mage lh.le lh.re rh.le rh.re      sex     yi     vi
1      E.1.d   NA    93    17   130   760 combined 3.4384 0.0768
2      E.1.b   37   140    91   305   697 combined 1.2544 0.0228
5   E.1.a.ii   22    16    14    43   114 combined 1.0970 0.1613
6      E.1.b   25   102    97   597  1898 combined 1.2061 0.0222
7   E.1.a.ii   20    17    14    38    80 combined 0.9257 0.1645
8   E.1.a.ii   20    10     2    52   170 combined 2.6130 0.5202
```

The computed effect size (`"yi"`) and variance (`"vi"`) are added to the data set. The data set provides information about the following moderators: the publication year (`"year"`), whether the selection of subjects was based on eye-dominance or handedness (`"selection"`, with two categories), the type of investigator (`"investigator"`, with three categories[7]), the method to assess handedness (`"hand_assess"`, with two categories), the methods to assess eye-dominance (`"eye_assess"`, with six categories), the average age (`"mage"`).

[7]The three categories are "psychologist", "educationalist", "and other"

Both meta-regression and meta-CART are applied to re-analyze this data set. The analyses results are compared to illustrate the different research questions that can be answered by these approaches. For both analyses, the moderators "selection", "investigator", "hand_assess", and "eye_assess" are selected. The random effects model is chosen and the DerSimonian-Laird estimator is used to estimate the residual heterogeneity. We use "`set.seed(2018)`" so that the analysis results described in this paper can be replicated.

First, to answer the question "which moderators affect hand-eye association?", we fit a meta-regression model with the main effects of the selected moderators.

```
R> set.seed(2018)
R> regHAND.re <- rma(yi = yi, vi = vi, method = "DL",
+                 mods = ~ eye_assess + investigator + hand_assess +
+                  selection, data = dat.handedness)
R> regHAND.re

Mixed-Effects Model (k = 54; tau^2 estimator: DL)

tau^2 (estimated amount of residual heterogeneity):     0.1268 (SE = 0.0726)
tau (square root of estimated tau^2 value):             0.3561
I^2 (residual heterogeneity / unaccounted variability): 60.39%
H^2 (unaccounted variability / sampling variability):   2.52
R^2 (amount of heterogeneity accounted for):            58.26%

Test for Residual Heterogeneity:
QE(df = 44) = 111.0699, p-val < .0001

Test of Moderators (coefficient(s) 2:10):
QM(df = 9) = 51.3955, p-val < .0001

Model Results:

                        estimate      se     zval    pval    ci.lb
intrcpt                   1.3564  0.4468   3.0356  0.0024   0.4806
eye_assessE.1.a.ii       -0.6513  0.3222  -2.0214  0.0432  -1.2828
eye_assessE.1.b          -0.9618  0.3585  -2.6829  0.0073  -1.6645
eye_assessE.1.c          -0.8967  0.3486  -2.5718  0.0101  -1.5800
eye_assessE.1.d           0.6733  0.4381   1.5369  0.1243  -0.1854
eye_assessE.2.a           0.2246  0.3898   0.5762  0.5645  -0.5394
investigatorother         0.4302  0.3774   1.1400  0.2543  -0.3094
```

```
investigatorpsychologist    0.4222  0.3510   1.2027  0.2291  -0.2658
hand_assessquestionnaire    0.0595  0.1913   0.3108  0.7559  -0.3156
selectionyes                0.0117  0.2392   0.0491  0.9609  -0.4571
                            ci.ub
intrcpt                    2.2322  **
eye_assessE.1.a.ii        -0.0198   *
eye_assessE.1.b           -0.2592  **
eye_assessE.1.c           -0.2133   *
eye_assessE.1.d            1.5320
eye_assessE.2.a            0.9887
investigatorother          1.1698
investigatorpsychologist   1.1101
hand_assessquestionnaire   0.4345
selectionyes               0.4806

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that three contrasts derived from the method to assess eye-dominance are significant: contrasted with the eye-dominance assessment method "E.1.a.i" (i.e., the reference method), "E.1.a.ii", "E.1.b", and "E.1.c" result in significantly lower hand-eye association. The remaining two methods, "E.2.a" and "E.1.d" do not differ significantly from the reference method. Note that the interpretation of these contrasts depends highly on the choice of reference method.

Further analyses are needed if we are interested in whether some categories of a moderator can be combined and whether interaction effects between the moderators are present. Alternatively, meta-CART can be performed to answer the question "which combinations of (categories of) moderators are influential?". The combinations can be either combinations of multiple moderators, or combinations of multiple categories of one moderator. We fit a RE meta-CART model by

```
R> library(metacart)
Loading required package: rpart
Loading required package: ggplot2
Loading required package: gridExtra
R> modHAND.re <- REmrt(yi ~ eye_assess + investigator + hand_assess
+                     + selection, data = dat.handedness,  vi = vi,
+                     c = 0.5, maxL = 10L, minsplit = 10L, cp = 1e-04,
+                     minbucket = 3, xval = 10, lookahead = FALSE)
R> summary(modHAND.re)
```

```
Random Effects meta-tree (K = 54 studies);
metacartv2::REmrt(formula = yi ~ eye_assess + investigator +
hand_assess + selection, data = dat.handedness, vi = vi,
c = 0.5, maxL = 10L, minsplit = 10L, cp = 1e-04, minbucket = 3,
lookahead = FALSE, xval = 10)

A tree with 2 terminal nodes was detected
Moderators were detected as: eye_assess

Test for Between-Subgroups Heterogeneity under RE assumption:
Qb = 46.293 (df = 1), p-value < 1e-04;
The estimate for the residual heterogeneity tau2 = 0.120;

Subgroup Meta-analysis Results:
   K     g    se   zval  pval ci.lb ci.ub
2 39 0.988 0.086 11.553 0.000 0.821 1.156 ***
3 15 2.071 0.134 15.438 0.000 1.808 2.334 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The `plot` function can be used to inspect the influence of eye-dominance assessment methods on the association between handedness and eye-dominance.

```
R> plot(modHAND.re)
```

The plot can be found in Figure 4.1, which shows a tree with one split and two terminal nodes. Note that this tree results from pruning an initially large tree based on cross-validation (see 4.2.2). According to the summary results and the plot, the moderator "eye_assess" has a strong influence on the association between handedness and eye-dominance ($Q_B^* = 46.29$, df = 1, p-value < 0.0001). When the eye-dominance is assessed using the methods "E.1.a.ii", "E.1.b", or "E.1.c", the observed association between handedness and eye-dominance is lower (log odds-ratio = 0.988, 95% CI: 0.821 to 1.156). When the eye-dominance is assessed using the methods "E.1.a.i", "E.1.d" or "E.2.a", the observed association is generally higher (log odds-ratio = 2.071, 95% CI: 1.808 to 2.334). This is in accordance with the claims in Bourassa et al. (1996) that the methods to assess eye-dominance can be partitioned into two subgroups: 1) unbiased methods including "E.1.a.ii: monocular procedure with object/instrument held in both hands", "E.1.b: binocular procedure", and "E.1.c: a combination of the previous methods", 2) biased methods including "E.1.a.i: monocular procedure with object/instrument held in one hand", "E.2.a: assessment based on a questionnaire", and "E.1.d: some other method". The methods "E.1.a.ii", "E.1.b", or "E.1.c" were symmetric in the sense that

both hands are used equivalently during the process of measurement, and therefore are less likely to have measurement bias.

Comparing to meta-regression, which estimates the coefficients and tests the significance for the contrasts derived from "eye_assess", meta-CART partitions the multi-categorical moderator "eye_assess" into subgroups, and tests the heterogeneity between the resulting subgroups. The subgrouping membership can be used to verify prior hypotheses[8], or to generate hypotheses if no a priori hypotheses exist.

4.4.2 Identify interaction effects using FE/RE assumptions

In this example, we perform a meta-analysis to identify the most effective combination of treatment components by exploring the interactions between them. The **metacart** package provides the data object `dat.BCT2009` as a subset of the meta-analytic data from Michie, Abraham, et al. (2009). The meta-analysis by Michie, Abraham, et al. (2009) aimed to assess the effectiveness of interventions designed to promote physical activity and health eating, and investigated whether theoretically specified behavior change techniques (BCTs) improve the effectiveness. The subset used in this example consists of 106 interventions that included at least one of the motivation-enhancing BCTs.

```
R> data("dat.BCT2009")
R> head(dat.BCT2009, n = 5)
```

```
                study       g          vi T1  T2  T3  T4  T25
1   ALDANA 2005 (PA) 0.60751 0.02469612  1   1   0   1    0
2      ANDERSON 2006 0.74651 0.05116192  0   0   0   1    0
3     ARAO 2007 (PA) 0.73090 0.06694639  0   0   0   1    0
4 BABAZONO 2007 (PA) 0.88197 0.09911163  0   0   0   1    0
5          BAKER 2008 0.73575 0.05308416  0   1   0   0    0
```

```
R> summary(dat.BCT2009)
```

```
                study           g                  vi          T1
ALDANA 2005 (PA)  :  1   Min.   :-0.1696   Min.   :0.00129   0:69
ANDERSON 2006     :  1   1st Qu.: 0.1291   1st Qu.:0.01533   1:37
ARAO 2007 (PA)    :  1   Median : 0.3033   Median :0.02796
BABAZONO 2007 (PA):  1   Mean   : 0.3308   Mean   :0.04985
BAKER 2008        :  1   3rd Qu.: 0.4932   3rd Qu.:0.06188
BENNETT 2008      :  1   Max.   : 1.2831   Max.   :0.45788
(Other)           :100
```

[8] If the subgroup membership is tested as a categorical moderator, meta-regression yields the same results as meta-CART.

```
 T2      T3     T4    T25
0:42   0:106   0:33   0:89
1:64           1:73   1:17
```

As displayed above, the data set contains information about the name and the publication year, the estimated effect size (g = Hedges' g, which denotes the standardized mean difference in treatment outcome), the sampling variance of the effect size, and whether specific BCTs were applied or not in a study ("0" for absent and "1" for present). In this example, we will re-analyze this data set focusing on the motivation-enhancing BCTs that may explain the heterogeneity in the effect sizes of the interventions. We use the pruning parameter $c = 0.5$ for both FE and RE meta-CART analyses (for more guidelines about the value of the pruning parameter c, see Li et al. 2019). Four moderators "T1: Provide information about behavior-health link", "T2: Provide information on consequences", "T4: Prompt intention formation", and "T25: Motivational interviewing" are included in the meta-CART analysis. The moderator "T3: Provide information about other's approval" is excluded since none of the studies applied this BCT.

If we have prior knowledge that the FE assumption is reasonable here, a FE meta-CART model can be fitted using the following code

```
R> modBCT.fe <- FEmrt(g ~ T1 + T2 + T4 + T25, data = dat.BCT2009,
+                  vi = vi, c = 0.5, control = rpart.control(xval = 10,
+                  minbucket = 5, minsplit = 10, cp = 1e-04))
```

To obtain the summary output, we use

```
R> summary(modBCT.fe)

Fixed Effects meta-tree (K = 106 studies);
FEmrt(formula = g ~ T1 + T2 + T4 + T25, data = dat.BCT2009, vi = vi,
c = 0.5, control = rpart.control(xval = 10, minbucket = 5,
minsplit = 10, cp = 1e-04))

A tree with 3 terminal nodes was detected
Moderators were detected as: T1, T4

Test for Between-Subgroups Heterogeneity under FE assumption:
Qb = 40.589 (df = 2), p-value < 1e-04;

Subgroup Meta-analysis Results:
   K      Qw     g    se   zval  pval ci.lb ci.ub
2 69 168.397 0.202 0.014 14.521 0.000 0.175 0.229 ***
```

```
6 15  20.195 0.191 0.037  5.115 0.000 0.118 0.264 ***
7 22  23.855 0.438 0.035 12.534 0.000 0.369 0.506 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

This output shows that a tree with three terminal nodes was detected. The studies are partitioned into three subgroups based on two influential moderators "T1" and "T4". The between-subgroups heterogeneity is significant ($Q_B = 40.589$, df $= 2$, p-value < 0.0001). The subgroup analysis results show the number of studies (K), the within-subgroup Q-statistic, the summary effect sizes (g) in each subgroup, the standard errors of the summary effect sizes (se), Z-test statistics of the summary effect sizes, and the confidence intervals of the summary effect sizes. The `plot` function can be used to inspect the final tree, in this case, the interaction between the two moderators "T1" and "T4".

```
R> plot(modBCT.fe)
```

The plot is shown in Figure 4.3. If an intervention does not include "T1" (i.e., $T_1 = 0$ is true), then the intervention ends up in the left terminal node (with $K = 69$). For those that include "T1" but not include "T4", they end up in the middle terminal node (with $K = 15$). Interventions including both "T1" and "T4" end up in the right terminal node (with $K = 22$). Combined with the other results, we can see that the summary effect size is the highest when "T1" and "T4" are both present in the intervention ($g = 0.438$, 95% CI: 0.369 to 0.506).

If a new intervention is designed with three BCTs "T1", "T2" and "T25" (thus without "T4"), the prediction of its effect size based on this tree (`modBCT.fe`) can be obtained by the `predict` function.

```
R> newStudy <- data.frame(T1 = factor(1, levels = 0:1),
+                 T2 = factor(1, levels = 0:1),
+                 T4 = factor(0, levels = 0:1),
+                 T25 = factor(1, levels = 0:1))
R> predict(modBCT.fe, newStudy)

          g     ci.lb    ci.ub T1 T2 T4 T25
1 0.1906789 0.1176188 0.263739  1  1  0   1
```

When interpreting the FE meta-CART analysis results, it is important to realize that the FE assumption ignores the uncertainty introduced by the residual heterogeneity. As a result, the confidence intervals of the summary effect sizes are more narrow than those estimated by using the RE model.

If we would like to take into account the residual heterogeneity, a RE meta-CART model can be fitted using the following commands

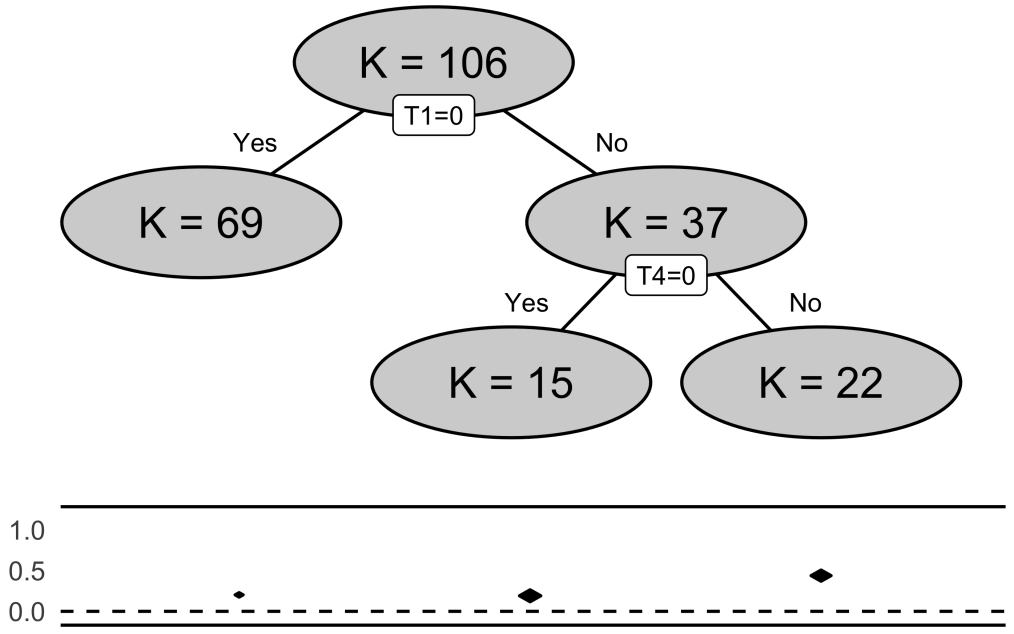


Figure 4.3: The meta-CART analysis result of 106 studies that examine the influence of motivation-enhancing BCTs on healthy eating and physical activities. The figure shows the FE meta-CART structure with splitting information at each internal node and the number of studies in each subgroup implied by a terminal node. The two solid lines show the range of the effect sizes of all the studies. The diamonds between the solid lines present the 95% confidence intervals of the summary effect sizes.

```
R> modBCT.re <- REmrt(g ~ T1 + T2 + T4 + T25, data = dat.BCT2009, vi = vi,
+                  c = 0.5, maxL = 10L, minsplit = 10L, cp = 1e-04,
+                  minbucket = 3, xval = 10, lookahead = FALSE)
Warning message:
In REmrt(g ~ T1 + T2 + T4 + T25, data = dat.BCT2009, vi = vi, c = 0.5,  :
no moderator effect was detected
R> summary(modBCT.re)

Random Effects meta-tree (K = 106 studies);
REmrt(formula = g ~ T1 + T2 + T4 + T25, data = dat.BCT2009, vi = vi,
c = 0.5, maxL = 10L, minsplit = 10L, cp = 1e-04, minbucket = 3,
xval = 10, lookahead = FALSE)

No moderator effect was detected

Test for Heterogeneity
Q = 253.0357 (df = 105), p-value < 1e-04;
The estimate for the residual heterogeneity tau2 = 0.023;

Random Effects Meta-analysis Results:
K      g     se   zval   pval  ci.lb  ci.ub
106  0.270  0.022 12.344  0.000  0.227  0.313    ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

The output shows that no influential moderators were identified. In this case, the `summary` method shows the standard RE meta-analysis results instead of the subgroup analysis results. The heterogeneity among the studies is significant ($Q_B^* = 253.036$, df $= 105$, $p-$value < 0.0001). The estimated summary effect size for all studies ($K = 106$) is 0.270 (95% CI: 0.227 to 0.313).

4.4.3 Look-ahead strategy

The **metacart** package provides a simulated data set `dat.balanced` to illustrate the look-ahead strategy for RE meta-CART analysis. The simulated data set contains $K = 60$ studies with four moderators: x_1, x_2, x_3, x_4, among which x_1, x_2, and x_4 are randomly sampled dichotomous variables and x_3 is sampled from a uniform distribution $U(0, 1)$. The sample size n was generated from a normal distribution $\mathcal{N}(160, (\frac{160}{3})^2)$. The residual heterogeneity was set as $\sigma_\tau^2 = 0.01$. The true model used to generate data was

$$d_k = 0.5 \cdot I(x_1 = 0, x_2 = 1) + 0.5 \cdot I(x_2 = 0, x_1 = 1) + \tau_k + \epsilon_k. \tag{4.9}$$

This model is similar to the simulated example used by Tibshirani and Knight (1999). This model is shown by Tibshirani and Knight (1999) to be difficult for greedy search procedures like CART because there is no information on where to split at the top level. Due to the same reason, standard meta-CART is likely to end up with a local optimum solution in such case. In this example, we would like to show that the look-ahead strategy described in Section 4.2.2 can alleviate this problem.

First, we inspect the data and fit a RE meta-CART model without using the look-ahead strategy.

```
R> data("dat.balanced")
R> head(dat.balanced)
```

```
            g          vi  x1 x2        x3 x4
1  0.08365950 0.02791743  0  0 0.03468667  1
2 -0.07713184 0.01252533  0  0 0.66439264  1
3  0.65561074 0.07416912  0  1 0.01766560  1
4  0.28081322 0.02305076  0  1 0.29408049  1
5  0.29945301 0.01215463  0  1 0.42494663  1
6  0.02193797 0.01452387  1  1 0.68668759  1
```

```
R> res <- REmrt(g ~ x1 + x2 + x3 + x4, vi = vi, data = dat.balanced,
+                         c = 1, maxL = 5L, minsplit = 6L, cp = 1e-04,
+                         minbucket = 3L, xval = 10, lookahead = FALSE)
Warning message:
In REmrt(g ~ x1 + x2 + x3 + x4, vi = vi, data = dat.balanced,  :
no moderator effect was detected
```

The output shows that meta-CART failed to detect any moderator. The reason is that the algorithm got stuck at a local optimum in the splitting procedure. This can be verified by inspecting the initial tree, unpruned by cross-validation.

```
R> res$initial.tree
```

```
Qb       tau2      split  mod pnode
1  0.000000 0.09243657       <NA> <NA>    NA
2  3.040267 0.08828453  x3 < 0.91   x3     1
3  7.317024 0.08224968  x3 < 0.27   x3     2
4 18.303908 0.06847671 x3 < 0.072   x3     4
5 22.521824 0.06555919 x3 < 0.035   x3     6
6 24.970538 0.06449154     x4 = 1   x4     3
```

This output shows the between-subgroups Q-statistic, the residual heterogeneity, the chosen split point (by default the values are shown in two decimal places), the chosen moderator, and the parent node to be partitioned at each split, with the first row presenting the root node when no split occurs. From the output we can see that the algorithm falsely chose the first splits at x_3 and ended up with $Q_B = 24.97$ after five splits.

Then we fit a RE meta-CART model using the look-ahead strategy.

```
R> res2 <- REmrt(g ~ x1 + x2 + x3 + x4, vi = vi, data = dat.balanced,
+                              c = 1, maxL = 5L, minsplit = 6L, cp = 1e-04,
+                              minbucket = 3L, xval = 10, lookahead = TRUE)
R> summary(res2)

Random Effects meta-tree (K = 60 studies);
metacartv2::REmrt(formula = g ~ x1 + x2 + x3 + x4, data = dat.balanced,
vi = vi, c = 1, maxL = 5L, minsplit = 6L, minbucket = 3L,
lookahead = T, cp = 1e-04, xval = 10)

A tree with 4 terminal nodes was detected
Moderators were detected as: x1, x2

Test for Between-Subgroups Heterogeneity under RE assumption:
Qb = 145.004 (df = 3), p-value < 1e-04;
The estimate for the residual heterogeneity tau2 = 0.017;

Subgroup Meta-analysis Results:
   K      g    se   zval  pval  ci.lb ci.ub
4 16 -0.054 0.044 -1.244 0.214 -0.140 0.031
5 17  0.516 0.045 11.413 0.000  0.427 0.605 ***
6 12 -0.017 0.051 -0.334 0.738 -0.117 0.083
7 15  0.518 0.046 11.358 0.000  0.429 0.607 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> plot(res2)
```

The results show that the RE meta-CART model with look-ahead strategy successfully recovered the true model as in (4.9). A tree with four terminal nodes was detected. As shown in Figure 4.4, the terminal node with $x_1 = 0$ and $x_2 = 1$ ($K = 17$) and the terminal node with $x_1 = 1$ and $x_2 = 0$ ($K = 15$) have higher effect sizes with CIs (presented by the diamonds) covering 0.5, whereas the terminal node with x_1 and x_2 both equal to 0 ($K = 16$) and the terminal node with x_1 and x_2 both equal to 1 ($K = 12$) have lower effect sizes with CIs covering 0.

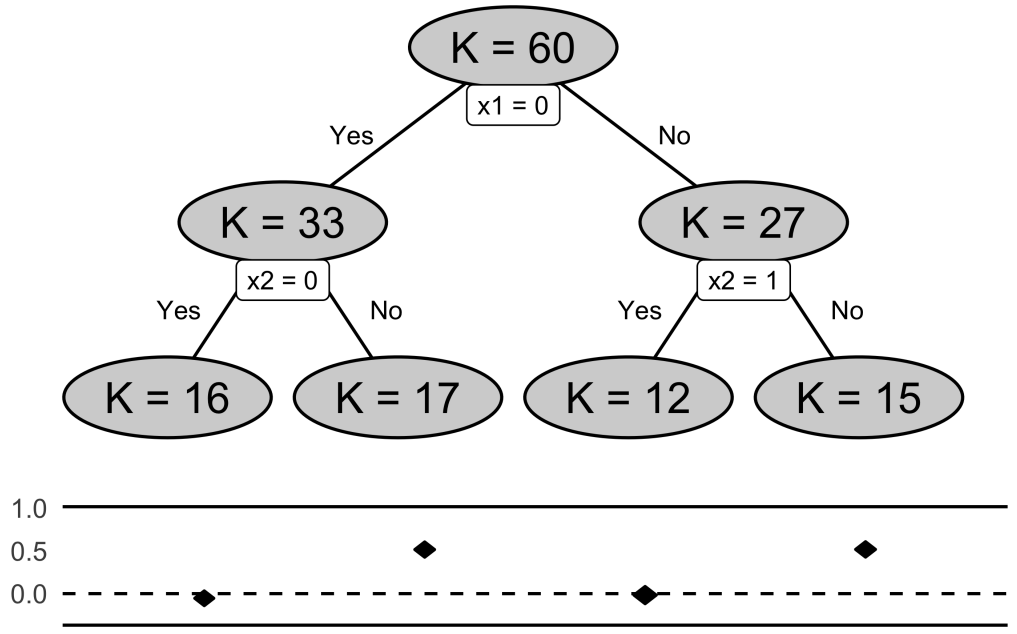


Figure 4.4: The analysis results of simulated data generated by fitting a RE meta-CART model with look-ahead strategy. The analysis successfully recovered the true model that was used to generate the simulated data.

The initial tree obtained with look-ahead strategy can be inspected by

```
R> res2$inital.tree
```

```
            Qb       tau2      split  mod pnode
1   0.0000000 0.092436572       <NA> <NA>    NA
2   0.2548616 0.092830849    x1 = 0   x1     1
3  42.0137849 0.049149501    x2 = 0   x2     2
4 145.0037965 0.016652079    x2 = 1   x2     3
5 188.0117000 0.011161345 x3 < 0.26   x3     6
6 213.7011113 0.008802224  x3 < 0.6   x3     4
```

Comparing this tree to the initial tree obtained from the previous model without the look-ahead strategy, the look-ahead strategy correctly finds the first two splits, and obtains a solution with much larger between-subgroups Q-statistic ($Q_B = 213.70$ after five splits).

4.5 Conclusions

This paper presents the **metacart** package written in R to perform meta-CART analysis both for fixed effect and random effects models. The algorithms and the main functions of the package **metacart** are described in sections 2 and 3, respectively. Applications

of the **R**-package **metacart** were illustrated through example analyses on two real-world data sets and one simulated data set in section 4.

One strength of the package **metacart** is that it can easily explore the interaction effects among multiple moderators using an interpretable tree model. Furthermore, for multi-categorical variables, it creates automatically the contrasts between (combinations of) categories that account for the highest amount of heterogeneity. Another strength is that **metacart** provides researchers various options for the tree growing and pruning processes. For example, researchers can choose the minimum number of studies in parent nodes and the minimum number of studies in terminal nodes based on the total sample size. In general, it is recommended to have at least three studies in each terminal nodes. The pruning parameter can be chosen based on the balance between power and Type I error. A detailed guideline for choosing the pruning parameter can be found in Li et al. (2019). The look-ahead strategy is recommended to partially relieve the local optimization problem when fitting a RE meta-CART model.

In conclusion, this paper and the developed **metacart** package introduce researchers to the implementation of meta-CART analysis, to facilitate exploring interaction effects between multiple moderators in the framework of meta-analysis.

Chapter 5

Interventions to promote healthy eating, physical activity and smoking in low-income groups: a systematic review with meta-analysis of behavior change techniques and delivery/context

abstract

Purpose: Healthy eating, physical activity and smoking interventions for low-income groups may have small, positive effects. Identifying effective intervention components could guide intervention development. This study investigated which content and delivery components of interventions were associated with increased healthy behavior in randomised controlled trials (RCTs) for low-income adults.

Method: Data from a review showing intervention effects in 35 RCTs containing 45 interventions with 17,000 participants were analysed to assess associations with behavior change techniques (BCTs) and delivery/context components from the template for intervention description and replication (TIDieR) checklist. The associations of 46 BCTs and 14 delivery/context components with behavior change (measures of healthy eating, physical activity and smoking cessation) were examined using random effects subgroup meta-analyses. Synergistic effects of components were examined using classification and regression trees (meta-CART) analyses based on both fixed and random effects assumptions.

Results:

For healthy eating, self-monitoring, delivery through personal contact, and targeting multiple behaviors were associated with increased effectiveness. Providing feedback, information about emotional consequences, or using prompts and cues were associated with reduced effectiveness. In synergistic analyses, interventions were most effective without feedback, or with self-monitoring excluding feedback. More effective physical activity interventions included behavioral practice/rehearsal or instruction, focussed solely on physical activity or took place in home/community settings. Information about antecedents was associated with reduced effectiveness. In synergistic analyses, interventions were most effective in home/community settings with instruction. No associations were identified for smoking.

Conclusion:

This study identified BCTs and delivery/context components, individually and synergistically, linked to increased and reduced effectiveness of healthy eating and physical activity interventions. The identified components should be subject to further experimental study to help inform the development effective behavior change interventions for low-income groups to reduce health inequalities.

5.1 Background

People of lower socioeconomic status are less likely to eat healthily (Drewnowski & Specter, 2004) or be physically active and (Stamatakis, 2006) more likely to smoke (Government, 2008) compared to those of higher socioeconomic status. These behaviors may be mediators of the well-established link between social position and morbidity and mortality outcomes (Gruer, Hart, Gordon, & Watt, 2009; Hart, Gruer, & Watt, 2011; Whitley, Batty, Hunt, Popham, & Benzeval, 2013). Amongst many socioeconomic indicators including education levels, measures of job status and access to healthcare, personal or household income is a direct economic indicator which is strongly positively correlated with health outcomes (Drewnowski & Specter, 2004). In trials of interventions for the general population, people with a lower income may experience poorer behavior change outcomes than more affluent participants potentially leading to intervention-generated inequalities (Chesterman, Judge, Bauld, & Ferguson, 2005; Hiscock, Judge, & Bauld, 2010; Niederdeppe, Fiore, Baker, & Smith, 2008; White, Adams, & Heywood, n.d.). Targeting health promotion efforts at people facing deprivation may prevent ill health and contribute towards reducing health inequalities (Gruer et al., 2009).

A previous review of interventions targeted at low-income participants found that approximately half were effective (Michie, Jochelson, Markham, & Bridle, 2009). Furthermore, a more recent systematic review with meta-analysis found positive, but small and variable effects on healthy eating (standardised mean difference (SMD) 0.22, $I^2 = 48\%$, physical activity (SMD 0.21, $I^2 = 76\%$) and smoking (relative risk (RR) 1.59, $I^2 = 40\%$), smaller than other similar interventions with participants of mixed income (Bull, Dombrowski, McCleary, & Johnston, 2014). Initial explorations of heterogeneity were conducted in that review but associations between specific intervention components with variation in intervention effectiveness were not examined. Understanding this variability, including identifying potentially underutilised effective components with these groups, is important when health inequalities continue to widen (White et al., n.d.).

Behavioral medicine researchers have recently developed several frameworks and tools to help us accurately and comprehensively describe intervention components and accumulate evidence of 'what works'. The template for intervention description and replication (TIDieR) checklist specifies 12 elements of healthcare interventions which study authors should report, including aspects of delivery and context (Hoffmann et al., 2014). These include describing 'how' they took place, i.e. the mode of delivery (e.g. face-to-face, telephone), 'where', or the setting (e.g. at home, in a school, or in a health facility) and 'what' content was delivered. For further characterising this, researchers have developed a shared language known as the behavior change technique taxonomy (BCTTv1) (Michie et al., 2013), including 93 active ingredients of behavior change interventions called behavior change techniques (BCTs). Better understanding the content, delivery and context of

existing behavior change interventions for low-income groups and exploring which seem effective and ineffective could prove timely and useful.

Recently, a promising new statistical method called Meta-CART has been developed to help analyse the effectiveness of combinations of BCTs and other intervention features (Dusseldorp et al., 2014; Li, Dusseldorp, & Meulman, 2017). Most health behavior change interventions are complex (Boutron, Moher, Altman, Schulz, & Ravaud, 2008) containing many BCTs and delivery/context components which can amplify or attenuate each other's effect (Dusseldorp et al., 2014). It has been argued that analyses must consider or control for this 'co-occurrence of methods' to advance behavior change science (Peters, Ruiter, & Kok, 2013). For instance, healthy eating interventions could involve combinations of goal setting, self-monitoring of behavior, and/or practical social support delivered by a health coach on the telephone or via a mailed leaflet. Traditional moderator analysis only examines the effect of each moderator individually, whereas meta-CART can use subgroup meta-analysis to identify interactions between moderators across interventions, such as to explore whether goal setting may be best delivered on the telephone or via a leaflet.

5.2 Aim and Objectives

This study aimed to conduct a new analysis of data from a previously published systematic review of health promotion interventions for low-income groups (Bull et al., 2014), applying behavioral science frameworks and new statistical methods to understand more about their effectiveness. While the previous paper found interventions to have small, positive effects, the current paper investigates which critical features of intervention content and delivery may contribute to their effectiveness. The association between a range of intervention components, individually and in combination, with variability in intervention effect sizes was examined. There were two specific objectives:

- To explore which individual BCTs and delivery/context features such as those from the TIDieR checklist are associated with effectiveness by applying moderator analyses.
- To explore synergistic effects between BCTs and delivery/context components and identify combinations associated with effectiveness by applying the new method meta-CART.

5.3 Methods

The study was registered in the PROSPERO database (CRD42015017468) and completed as per protocol. We applied moderator analyses to the data from a previously published systematic review with meta-analysis. For clarity, the original review's eligibility criteria,

search strategy and data collection processes are summarised below in this section, but further detail can be found in the published paper `http://bmjopen.bmj.com/content/4/11/e006046` (Bull et al., 2014).

5.4 Original Review Method Summary

The original review by Bull et al. (2014) included studies meeting the following inclusion criteria: (i) population: currently healthy adults described in the study as 'low-income'; (ii) interventions: aiming to change healthy eating, physical activity and/or smoking behavior in any combination; (iii) study design: RCTs or Cluster RCTs, with no limits on control condition design; (iv) outcomes: behavioral outcomes relevant to healthy eating, physical activity or smoking (e.g. self-reported portions of fruit per day, accelerometer-measured steps walked per week, or self-reported abstinence from smoking for seven consecutive days); (v) date: primary search carried out January 2006 to July 2014; (vi) language: English.

Bull et al. (2014) searched eight databases for studies with terms relating to low-income groups, terms for healthy eating, physical activity and smoking behaviors, and terms relating to interventions and health programs. In addition, in Bull et al. (2014) studies published between 1995 and 2006 were identified from another previously published review without meta-analysis on the topic (Michie, Jochelson, et al., 2009) rather than through a primary search and screened against the inclusion criteria listed above, since Michie, Jochelson, et al. (2009)used similar but broader search criteria which should have included all the relevant articles. Finally, in addition to these searches, Bull et al. (2014) checked each included study's bibliography for potentially relevant articles to screen.

In Bull et al. (2014), three authors screened titles and abstracts; one author screened full texts. In both stages, double screening of a random 10% yielded high inter-rater reliability. Data were collected using a piloted data extraction form based on Davidson et al. (2003). Three authors jointly extracted design, methods and results data. Proportions were extracted for dichotomous smoking outcomes; means and standard deviations were extracted for continuous healthy eating and physical activity outcomes. Where there was a choice, the outcome extracted was the primary measure specified by authors measured as objectively as possible, adjusted for baseline if the authors had thought this necessary. Risk of bias in individual studies was assessed based on standard criteria adapted from Avenell et al. (2004) and publication bias inspected visually using a funnel plot, reported in Bull et al. (2014).

5.5 Current Review Methods

In the new analysis, content and delivery/context component data were extracted from intervention descriptions in studies. Two authors jointly coded 14 components of each intervention, including 12 components based on the TIDieR checklist (Hoffmann et al., 2014) with the addition of 'WHO RECEIVED' the intervention and the outcome measure type (see Fig. 5.1 for the list of 14 components). A trained coder extracted each intervention's BCTs using BCTTv1 (Michie et al., 2013). Two expert coders extracted the BCTs in a random subset of 16 studies' intervention descriptions to assess coding reliability. Prevalence and bias-adjusted kappa (PABAK) (Byrt, Bishop, & Carlin, 1993) was 0.87 and 0.83 respectively, suggesting high inter-rater agreement (Orwin & Vevea, 2009). Published online supplementary materials were used where available and the corresponding author was contacted in the case of missing data.

1. **WHY: Theoretical base described?**	(Yes or No)
2. **WHAT: Number of behaviors targeted**	(One or more than one)
3. **HOW: Personal contact included?**	(Yes or No)
4. **HOW: Use of a manual, for those with personal contact?**	(Yes or No)
5. **HOW: Face-to-face component for those with personal contact?**	(Yes or No)
6. **HOW: Individual or group format, for those with personal contact?**	(Individual, group or both)
7. **WHO PROVIDED: Facilitator type, for those with personal contact**	(Professional people only; lay people only or both)
8. **WHO PROVIDED: Training specified, for those with personal contact?**	(Yes or No)
9. **WHERE: Intervention setting**	(Community; health or home setting)
10. **WHEN AND HOW MUCH: Intervention duration**	(Number of weeks)
11. **WHEN AND HOW MUCH: Intervention intensity**	(Number of inputs)
12. **WHEN AND HOW MUCH: Number of BCTs delivered**	(Number of BCTs)
13. **Outcome measure type**	(Self-report only or more objective measure reported)
14. **Who received intervention**	(Mixed sex or all women in study)

Figure 5.1: Fourteen delivery/context components based on the TIDieR checklist.

5.6 Statistical Analysis

5.6.1 Moderator Analysis

For continuous healthy eating and physical activity outcomes, standardised mean differences (SMDs) were calculated using Hedges g. For dichotomous smoking outcomes,

we calculated relative risk (RR) of smoking abstinence and applied the Cochran-Mantel-Haenszel test (Mantel & Haenszel, 1959). To minimise chance impact of single trials, we only examined BCTs and intervention delivery components identified as present or absent in at least three interventions as potential moderators, following Dombrowski et al. (2012). Each BCT and categorical delivery/context variable was examined testing subgroup differences using a mixed effects model (i.e., a random effects model for the within-subgroup effect sizes, and a fixed effect model for testing the between-subgroups heterogeneity, as recommended by Borenstein et al., 2009). Three continuous intervention components (intervention intensity, duration and number of BCTs) were analysed using meta-regression. We used comprehensive meta-analysis (v.2) for these analyses.

5.6.2 Meta-CART Analysis

To explore interaction effects and identify effective combinations of BCTs and delivery/context components, meta-CART was applied. Meta-CART is a tree-based method that combines the machine learning technique CART (Classification And Regression Trees) with meta-analysis (Breiman et al., 1984). Meta-CART uses study effect sizes (e.g. the study SMDs for healthy eating) as outcome variables, and potential moderators as predictor variables (e.g. BCTs). The method divides the study effect sizes into homogeneous subgroups of interventions based on influential moderators. The result of a meta-CART analysis is a tree, and the leaves (the end nodes) of the tree are subgroups of studies with similar combinations of moderators. In each subgroup, a pooled effect size is computed (i.e. a weighted average effect size). The pooled effect sizes of these subgroups are as different as possible since meta-CART maximises the between-subgroups heterogeneity (i.e. Q-statistic) (Li, Dusseldorp, & Meulman, 2017). At each split of the tree, the meta-CART algorithm searches for the moderator (i.e. a BCT/delivery or context variable) that maximises the between-subgroups differences. Initially, a large tree is grown with as many splits as possible. The best tree size is selected by 'pruning', removing the spurious splits of the tree based on the cross-validation error (Breiman et al., 1984). The final tree usually involves a smaller number of splits, for example a tree with two moderators (e.g. Fig. 5.3). The tree represents synergistic effects between the moderators involved in the splits of the tree. To test whether the synergistic effects significantly explain the heterogeneity between interventions, a new moderator variable is computed with categories referring to the end nodes (i.e. subgroups) of the tree, with each study belonging to one specific subgroup. Finally, a standard subgroup meta-analysis is performed using the new moderator variable to investigate whether the subgroup membership accounts for the heterogeneity in the study effect sizes.

Meta-CART analyses were performed using both random effects (RE) and fixed effect (FE) approaches to explore effective combination(s) of moderators. Both methods have

advantages: RE methods are more conservative, maximising control of type 1 error [16] whereas the more liberal FE method favours power, so applying both was seen as appropriate offering complementary information in this exploratory study. The RE meta-CART takes into account the residual heterogeneity unexplained by the individual BCTs and delivery/context components, whereas FE meta-CART assumes that the heterogeneity in study effect sizes is fully explained by the moderators. For both approaches, initial trees are grown with nodes of at least two interventions, before pruning. We used the half-standard-error pruning rule for the RE meta-CART analyses (i.e., selecting the smallest tree that has a cross-validation error within the minimum cross-validation error plus half times the standard error) [16]. The FE meta-CART used the minimum cross-validation error pruning rule, selecting the tree with minimum cross-validation error to increase power. Since the number of interventions was relatively small in our meta-analytic data sets, we performed leave-one-out cross-validation for pruning, although for larger data sets a tenfold cross-validation is generally recommended.

In total, six meta-CART analyses were performed: FE and RE analyses for healthy eating, physical activity and smoking interventions. Due to the relatively small number of interventions, for the meta-CART analyses we included the BCTs and delivery/context components which had shown significant effects in the univariate moderator analysis. All meta-CART analyses were performed using R (v.3.3.2).

5.7 Results

5.7.1 Original Review Study Selection and Characteristics

In the original review (Bull et al., 2014), 2569 titles and abstracts and 133 full texts were screened. Thirty-five trials were included with 17,000 adult participants with a low income. Thirty trials were conducted in the USA; three in the UK; one in Australia and Chile. Eleven studies recruited participants with a specific ethnic background (African-American, Latina or Chinese and Korean). The majority of participants were women (72.4%) living in the USA (77.2%) with a mean age of 38.6 years. The 35 studies contained 45 interventions: some studies targeted multiple behaviors or tested multiple interventions. In all, there were 16 interventions targeted at healthy eating; 12 at physical activity; 17 at smoking.

5.7.2 Current Review Study Characteristics

Of the 93 BCTs in the taxonomy and the 14 delivery/context components, 46 BCTs and all 14 delivery/context components were identified from the 45 published intervention descriptions. Each intervention contained between 2 and 20 BCTs (mean per intervention 6.62). Of the three delivery/context components which were continuous moderators

(WHEN AND HOW MUCH: Intervention duration, Intervention intensity and Number of BCTs delivered; Fig. 5.1), none were associated with effectiveness for any behavior, so here we present results for the categorical BCTs and 11 remaining delivery/context components only. Amongst these, four delivery/context components were not applicable to interventions without personal contact (use of a facilitator delivery manual; facilitator type; facilitator training; group or individual format) so were only examined for interventions with personal contact. In total, 23 BCTs and seven delivery/context components were present or absent in at least three interventions (see methods statistical analysis section) and analysed as potential moderators.

5.7.3 Healthy Eating: Individual Moderator Analysis

Sixteen BCTs and seven delivery/context components could be analysed within the 16 healthy eating interventions. Interventions including the BCT 2.3 Self-monitoring of behavior were associated with more healthy eating, while those with the BCTs 2.2 Feedback on behavior, 7.1 Prompts and cues or 5.6 Information about emotional consequences, were associated with less healthy eating (Supplementary Table 1). Amongst delivery/context components, including a face-to-face component (rather than remote contact, e.g. by telephone, or no personal contact) and a multi-behavioral focus (aiming to change both healthy eating and another behavior) were also associated with increased effectiveness (Supplementary Table 1). Figure 5.2 displays the statistically significant findings of the individual moderator analysis visually.

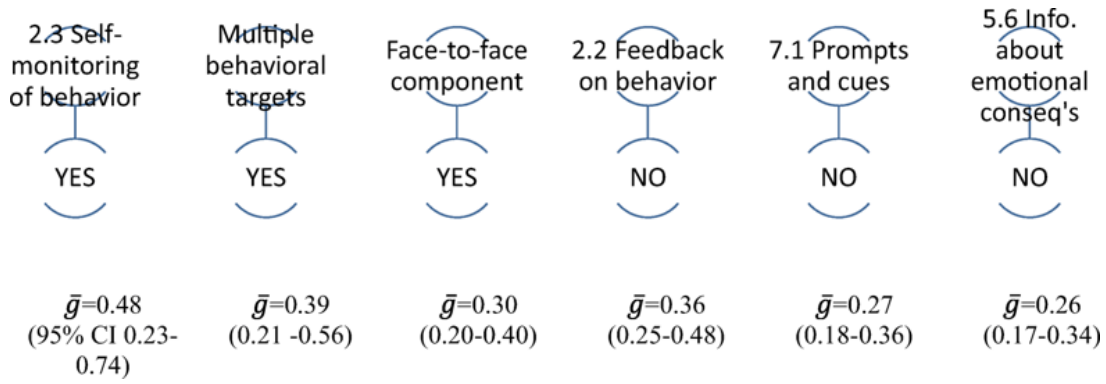


Figure 5.2: Diagram representing univariate moderator analyses for healthy eating. BCTs are presented with their original labels and number from BCTv1 (Michie et al., 2013). $\hat{g}$ represents effect size and 95% CIs statistical significance. This figure indicates that healthy eating interventions were significantly more effective where they did include the BCT 2.3 Self-monitoring of behavior, or if there were multiple behavioral targets or a face-to-face component, or did not include BCTs 2.2 Feedback on behavior, 7.1 Prompts and cues or 5.6 Information about emotional consequences.

5.7.4 Healthy Eating: Meta-CART Analysis of Synergistic Effects

Meta-CART was conducted to identify effective combinations of the four BCTs and two delivery/context components identified above in individual moderator analyses (Supplementary Table 1). The tree that resulted from the RE meta-CART analysis represented a synergistic effect between 2.2 Feedback on behavior and face-to-face component (Fig. 5.3). The interventions that excluded 2.2 Feedback on behavior showed the highest pooled effect size (i.e., $\bar{g} = 0.36$, 95% CI 0.26–0.46). When 2.2 Feedback on behavior was included, the interventions that also included a face-to-face component had a larger pooled effect size ($\bar{g} = 0.23$, 95% CI 0.14–0.31) than the interventions without ($\bar{g} = 0.10$, 95% CI 0.03–0.17). In the mixed effects subgroup meta-analysis, subgroups were significantly different from each other (between-subgroups Q- statistic = 17.49, $p = .002$).

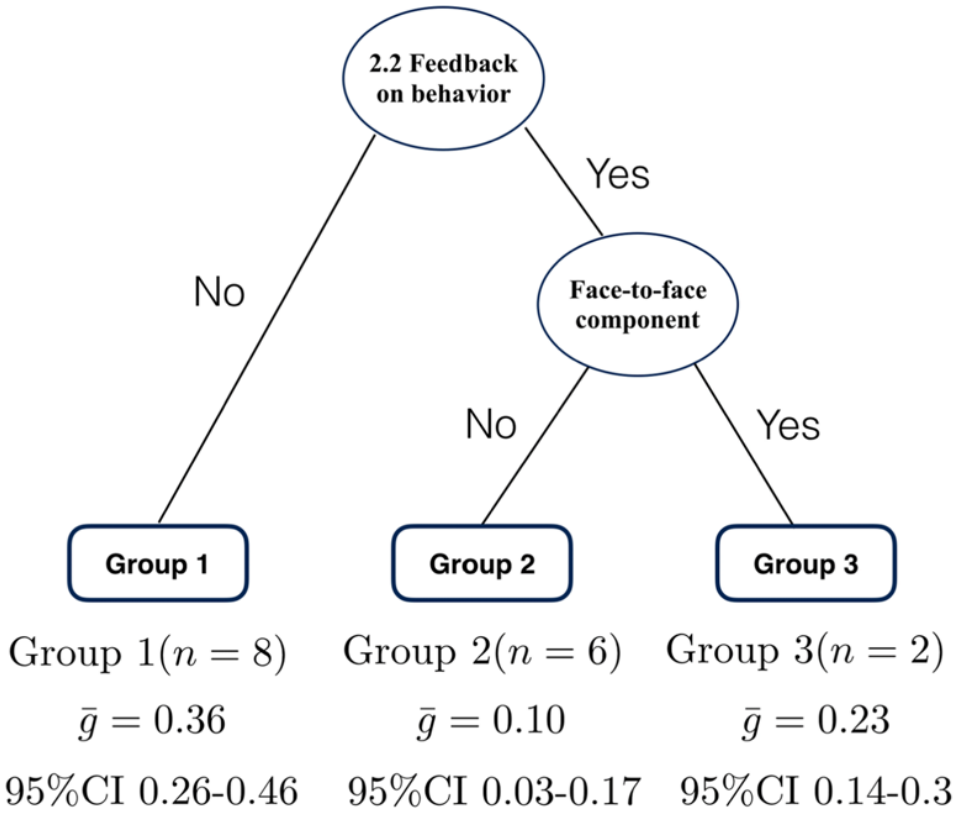


Figure 5.3: Results from random effects meta-CART meta-analysis for healthy eating ($K = 16$). This figure indicates random effects meta-CART analysis of effective combinations of the four BCTs and two delivery/context components identified as individually significant moderators in Fig. 5.2. Healthy eating interventions were more effective if they did not include the BCT 2.2 Feedback on behavior, but if they did, then those with a Face-to-face delivery component were more effective than those without.

Compared with the RE meta-CART analysis results, the tree resulting from the FE meta-CART analysis included one additional split: a synergistic effect between 2.2 Feedback on behavior and 2.3 Self-monitoring of behavior (Fig. 4). The interventions that

used 2.3 Self-monitoring of behavior but excluded 2.2 Feedback on behavior were most effective ($\bar{g} = 0.48$, 95% CI 0.29–0.66). The interventions using both 2.3 Self-monitoring of behavior and 2.2 Feedback on behavior were least effective ($\bar{g} = 0.31$, 95% CI 0.19–0.43). The synergistic effect between 2.2 Feedback on behavior and face-to- face component was the same as the RE meta-CART result. In the FE subgroup meta-analysis, again subgroups were significantly different from each other (between-subgroups Q-statistic = 19.68, $p = .002$).

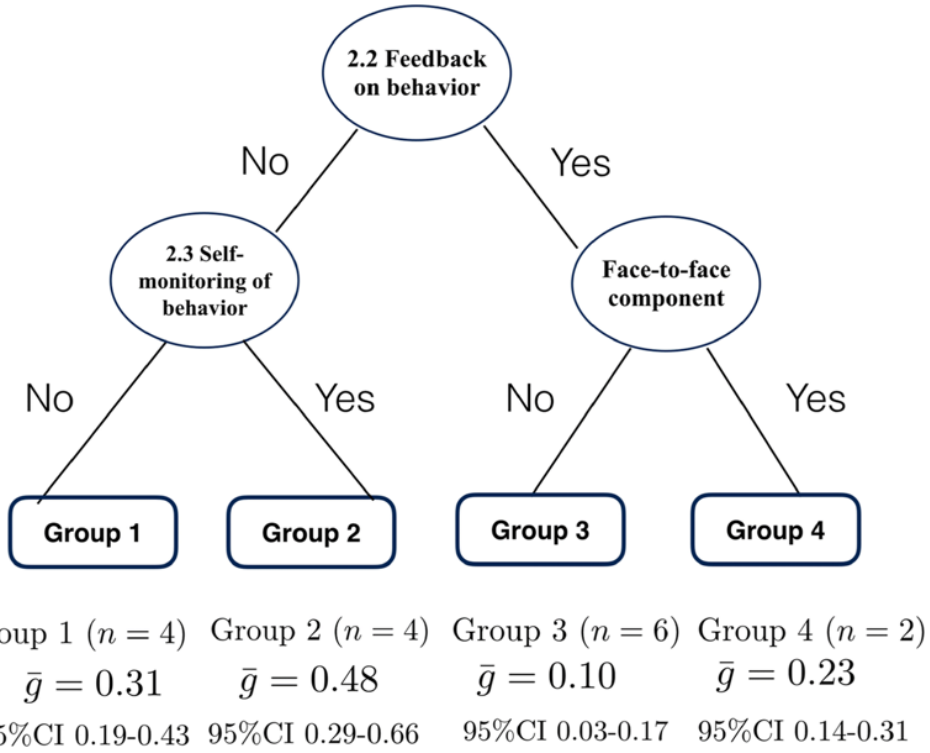


Figure 5.4: Results from fixed effects meta-CART meta-analysis for healthy eating ($K = 16$). This figure indicates fixed effects meta-CART analysis of effective combinations of the four BCTs and two delivery/context components identified as individually significant moderators in Fig. 5.2. Results were similar to Fig. 5.3, but also indicated that interventions excluding the BCT 2.2 Feedback on behavior but including 2.3 Self-monitoring of behavior were most effective.

5.7.5 Physical Activity: Individual Moderator Analysis

Fourteen BCTs and six delivery/context components could be analysed within the 12 physical activity interventions. Interventions including 8.1 Behavioral practice/rehearsal, or 4.1 Instruction on how to perform behavior) were associated with increased physical activity; interventions including the BCT 4.2 Information about antecedents with less physical activity. In addition, intervention delivery in a community or home setting (rather than in a health setting) and a sole focus on physical activity was associated with greater effectiveness (Supplementary Table 2). Figure 5.5 displays the statistically

significant findings visually of the individual moderator analysis.

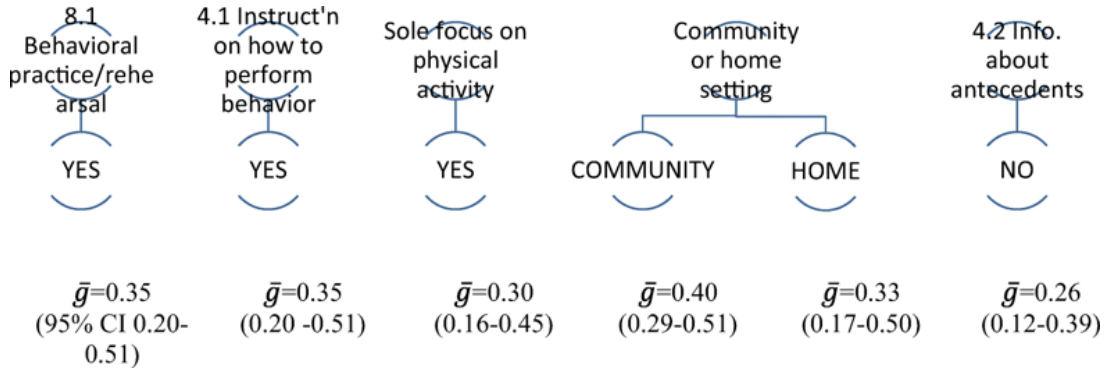


Figure 5.5: Diagram representing univariate moderator analyses for physical activity. This figure indicates that physical activity interventions were significantly more effective where they did include the BCTs 8.1 Behavioral practice/rehearsal or 4.1 Instruction on how to perform the behavior, or had a sole focus on physical activity, or were delivered in a community or home (rather than health) setting, or did not include the BCT 4.2 Information about antecedents.

5.7.6 Physical Activity: Meta-CART Analysis of Synergistic Effects

Meta-CART was conducted to identify effective combinations of the three BCTs and two delivery/context components identified as individual moderators above (Supplementary Table 2). RE meta-CART resulted in a tree with only one node: the root node: no combination of BCTs or delivery/context components was able to explain the heterogeneity in the effect sizes.

The tree resulting from the FE meta-CART analysis represented a synergistic effect between 4.1 Instruction on how to perform behavior and study setting. The interventions delivered in health settings had the lowest pooled effect size ($\bar{g} = -0.002$, 95% CI $-0.079 - 0.075$). Interventions delivered in community or home settings which included 4.1 Instruction on how to perform behavior had the highest pooled effect size ($\bar{g} = 0.42$, 95% CI 0.32–0.53). Interventions delivered in community or home settings but not including 4.1 Instruction on how to perform behavior had a pooled effect size in the middle of the other subgroups ($\bar{g} = 0.21$, 95% CI $0.02 - 0.40$). All of the interventions delivered in community or home settings applied either both 4.1 Instruction on how to perform behavior and 8.1 Behavioral practice /rehearsal or neither, none included just one. In the FE subgroup meta-analysis, subgroups were significantly different from each other (between-subgroups Q-statistic = 43.18, $p < .001$). Figure 5.6 displays this visually.

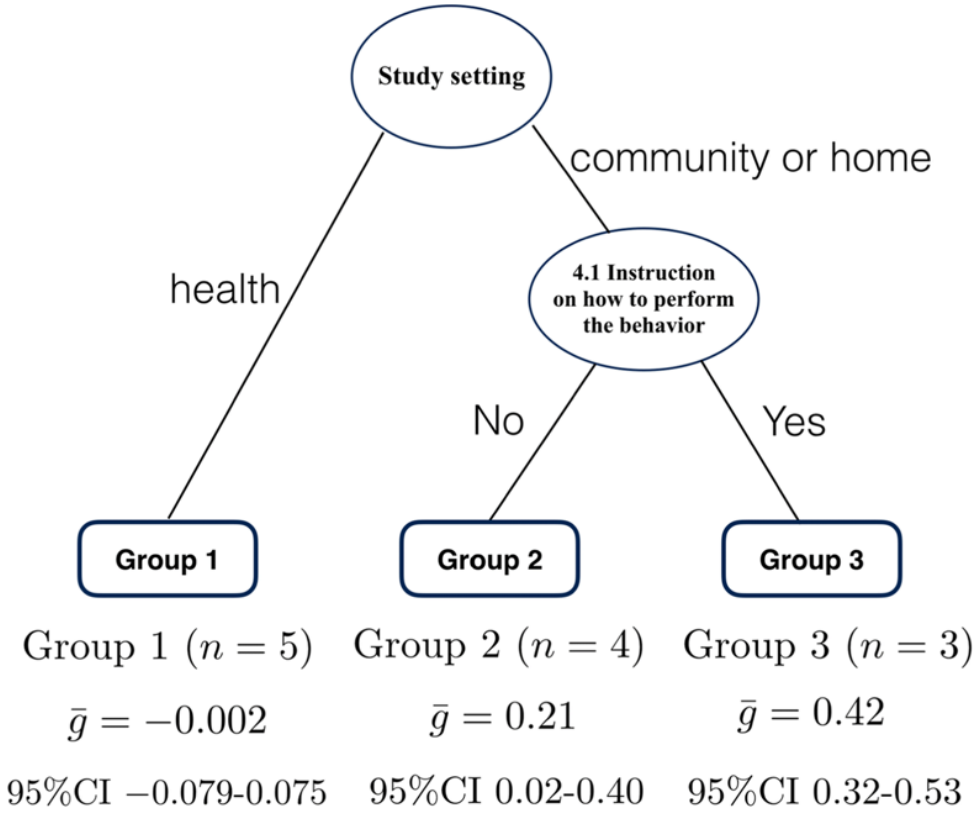


Figure 5.6: Results from fixed effects meta-CART meta-analysis for physical activity $K = 12$. This figure indicates fixed effects meta-CART analysis of effective combinations of the three BCTs and two delivery/context components identified as individually significant moderators in Fig. 5.5. Physical activity interventions were more effective if they were delivered in a community setting or at home and included the BCT 4.1 Instruction on how to perform the behavior, and were least effective if delivered in a health setting

5.7.7 Smoking: Individual Moderator Analysis

Eleven BCTs and five delivery/context components were analysed within the 17 smoking interventions, none of which were statistically associated with smoking intervention effectiveness (Supplementary Table 3).

5.7.8 Smoking: Meta-CART Analysis of Synergistic Effects

Both RE meta-CART and FE meta-CART detected no effective combination of BCTs or delivery/context components that could explain the heterogeneity in the effect sizes.

Table 5.1, summarises the individual BCTs and delivery/context components associated with increased or decreased effectiveness in healthy eating and physical activity with examples from the interventions included in the review.

Table 5.1: Definitions and examples of BCTs and delivery/context components associated with increased or decreased effectiveness.

	BCT or delivery/context component	Definition*BCT numbers, labels and definitions are taken from BCTV1 [14]	Example from interventions included in the review
Increased effectiveness			
DIET	2.3 Self-monitoring of behaviour*	Establish a method for the person to monitor and record their behaviour(s) as part of a behaviour change strategy*	In Keyserling et al. (2008), participants recorded daily fruit and vegetables consumed each day using a diary to help them increase this
	HOW: Face-to-face component included (yes)	For studies with personal contact, whether or not this personal contact was conducted face-to-face (instead of e.g. over telephone)	Emmons et al. (2005) included a counselling session in-person with a health advisor using motivational interviewing approaches to help support them to eat more healthily
	Number of behaviours targeted (multiple focus)	Whether the study aimed to change one behaviour (e.g. diet only) or multiple behaviours (e.g. diet and physical activity)	Jackson, Stotland, Caughey, and Gerbert (2011) focused on participants making healthy changes to both diet and physical activity
Decreased effectiveness			
DIET	2.2 Feedback on behaviour*	Monitor and provide informative or evaluative feedback on performance of the behaviour (e.g. form, frequency, duration, intensity)*	of Clinical Excellence. (2014) provided individualised written feedback to participants from an earlier assessment e.g. their current diet compared to national norms

	7.1 Prompts and cues*	Introduce or define environmental or social stimulus with the purpose of prompting or cueing the behaviour. The prompt or cue would normally occur at the time or place of performance*	Participants in Tessaro, Rye, Parker, Mangone, and McCrone (2007) received a portion magnet and wheel to put in their kitchen to remind them of healthy portion sizes
	5.6 Information about emotional consequences*	Provide information (e.g. written, verbal, visual) about emotional consequences of performing the behaviour*	Gans et al. (2009) included a video with testimonials from members of the target audience, who emphasised that eating healthier helps in feeling good about yourself and feeling better
Increased effectiveness			
PHYSICAL ACTIVITY	8.1 Behavioral practice/ rehearsal*	Prompt practice or rehearsal of the performance of the behaviour one or more times in a context or at a time when the performance may not be necessary, in order to increase habit and skill*	Marcus et al. (2013) included tailored written mailings which advised participants, for example, to make time for a ten minute walk one or two times each week, to help them build confidence that they can make exercise part of their weekly routine
	4.1 Instruction on how to perform a behavior*	Advise or agree on how to perform the behaviour (includes 'Skills training')*	Dangour et al. (2011)'s physical activity program for older adults included physical activity group training sessions where trained instructions guided participants in how to conduct activities e.g. warming up, chair stands, modified squats and arm pull-ups with rubber bands.

	WHERE: Study setting (community or at home, not in health setting)	Whether the study was set in the community, a health setting or at participants' home	Olvera et al. (2010)'s 12 week exercise program took place in community centres in the park, park playgrounds and grocery stores, as well as at school facilities, e.g. the school gym, playground or cafeteria
	Number of behaviors targeted (single focus)	Whether the study aimed to change one behavior (e.g. physical activity only) or multiple behaviours (e.g. diet and physical activity)	Dutton, Martin, Welsch, and Brantley (2007)'s intervention focused solely on increasing women's physical activity
Decreased effectiveness			
PHYSICAL ACTIVITY	4.2 Information about antecedents*	Provide information about antecedents (e.g. social and environmental situations and events, emotions, cognitions) that reliably predict performance of the behaviour*	Chang, Nitzke, and Brown (2010) provided examples of triggers relating to eating and being active in the environment as part of their behaviour change intervention

5.8 Discussion

In this study, we explored active components of interventions (BCTs) and the context and methods of delivery associated with effectiveness in health behavior change interventions for low-income adults, applying both individual moderator analyses and meta-CART to explore combinations of components. The content, context and delivery of effective interventions appeared different for healthy eating and physical activity behaviors. For healthy eating behavior, individual moderator analysis suggested that effective techniques could be to encourage self-monitoring, provide face-to-face contact with a facilitator or work

on physical activity simultaneously, without providing feedback on behavior, use prompts and cues or provide information about emotional consequences of healthy eating. Table 5.1 includes examples of each within the studies. These could substantially increase effectiveness: healthy eating interventions with self-monitoring had an SMD of 0.48 compared with the overall SMD of 0.22. Meta-CART analyses exploring combinations tended to confirm that interventions were more effective without feedback, especially when combined with self-monitoring, yet suggested that for interventions which did include feedback, then effects could be stronger when combined with face-to-face contact.

In individual moderator analyses, physical activity interventions tended to be more effective where they had a sole focus on participants being active or were delivered at home or in a community rather than health setting. Activity interventions were more effective where they included direct instruction or opportunities to practice and rehearse active movements but less effective when they included information about antecedents (see Table 5.1, for examples). Meta-CART suggested that a particularly effective combination could be instruction on how to perform the behavior within the community or home (not health) setting.

It may seem surprising that feedback on behavior was associated with lower healthy eating change, despite key behavior change guidance recommending it as a 'proven technique' (of Clinical Excellence., 2014). Yet the two healthy eating interventions with lowest effect sizes(Elder et al., 2006; Gans et al., 2009) provided feedback in a similar, non face-to-face way, through mailed written statements of previously reported healthy eating behavior, and meta-CART suggested that face-to-face delivery could mitigate against the lower effect size. In this review, mailed delivery formats were popular, and e-health technology is increasingly being employed to increase the reach of public health interventions (Norman et al., 2007) but our review suggested 'a personal touch' may be important to support low-income communities in their healthy eating efforts. The finding that self-monitoring without feedback was an effective combination was also surprising given that Control Theory (Carver & Scheier, 1982) would advocate their combination, and may oppose meta-regression findings that self-monitoring was more effective in healthy eating and physical activity interventions if combined with another Control Theory technique (Michie, Abraham, et al., 2009). Again this may be explained by other factors associated with studies where feedback was included.

The strong effect of instruction on physical activity behavior has been found in previous research (Williams & French, 2011), although this study adds that delivery of this BCT in a community or home rather than health setting is desirable. Thus, a behavior should be taught in the context that is likely to be (a) easy for participants to attend and (b) as similar as possible to real life to facilitate further performance as associating a behavior with the context is essential to building habits (Gardner, 2015). Additionally, recent evidence suggests that habit mediates the effects of planning on behaivour change

(Potthoff et al., 2017).

It may also seem counter-intuitive that interventions with a multi-behavior focus (targeting both healthy eating and physical activity) led to increased changes in healthy eating, but that in physical activity a single focus was preferable. Amongst several explanations, it could be assumed that most participants taking part in a multi-component intervention were aiming for weight loss. Since initial weight loss is more easily achieved by calorie restriction than by increased burning through exercise (Kushner, 2007), healthy eating may have been the core focus for both intervention facilitators and participants in these studies.

Another finding in this review was that the inclusion of information-focussed BCTs often used in public health interventions such as 4.2 Information about antecedents for physical activity and 5.6 Information about emotional consequences for healthy eating resulted in a less successful outcome. This builds on similar findings in different populations (Dombrowski et al., 2012), further evidence that information (particularly when directed at fear-arousing consequences) is likely to be ineffective without additional BCTs aimed at increasing self-efficacy or planning (Peters et al., 2013). A further possibility is that information-giving may dwarf the other more effective components; in a meta-analysis of combinations of components of internet-based interventions for the general public, van Genugten and colleagues (Van Genugten, Dusseldorp, Webb, & Van Empelen, 2016) found that interventions that were quick to deliver and easy to understand were more effective.

No BCTs or delivery/context variables were associated with smoking intervention effectiveness, in contrast to reviews with pregnant women and people with lung disease respectively which found positive effects for action planning amongst other BCTs(Bartlett, Sheeran, & Hawley, 2014; Lorencatto, West, & Michie, 2012). This may reflect the lower heterogeneity in smoking compared to healthy eating and physical activity effects in this review, study authors including a limited range of BCTs in interventions or perhaps poor intervention description (Lorencatto, West, Stavri, & Michie, 2012). We also found no association between theory use and effectiveness in this review, contrary to some reviews of behavioral interventions (Prestwich et al., 2014; Webb, Sniehotta, & Michie, 2010) but in line with recent diabetes intervention analyses [24, 53]. Similar to other reviews, employing a greater number of techniques was also not linked with increased effectiveness (Michie, Jochelson, et al., 2009; Webb et al., 2010).

Many BCTs and delivery/context components could not be analysed as they were seemingly rarely used (e.g. BCT 7.1: Prompts and cues, identified in one physical activity intervention) or used in all interventions: this could reflect poor reporting of behavioral intervention content (Hoffmann, Erueti, & Glasziou, 2013; McCleary, Duncan, Stewart, & Francis, 2013). Indeed in the smoking interventions, only 11 BCTs could be analysed. The smoking cessation field may be more extensively developed than others and so perhaps greater consensus has been reached on necessary components of stop smoking support

(Tobacco et al., 2008).

5.9 Supplementary material

Supplementary table 1: see https://static-content.springer.com/esm/art%3A10.1007%2Fs12529-018-9734-z/MediaObjects/12529_2018_9734_MOESM1_ESM.docx

Supplementary table 2: see https://static-content.springer.com/esm/art%3A10.1007%2Fs12529-018-9734-z/MediaObjects/12529_2018_9734_MOESM2_ESM.docx

Supplementary table 3: see https://static-content.springer.com/esm/art%3A10.1007%2Fs12529-018-9734-z/MediaObjects/12529_2018_9734_MOESM3_ESM.docx

Chapter 6

Advanced tree-based subgroup identification in meta-analysis

abstract

Background: In meta-analysis, heterogeneity often exists between studies. In such cases, it is essential to investigate the sources of heterogeneity and understand the relationship between effect size and study characteristics (i.e., moderators). Applying tree-based methods in meta-analysis is a promising alternative for conventional meta-regression, since trees excel at modeling interactions and non-linear relationships and provide easily interpretable results. In particular, a recently proposed method called meta-CART integrates classification and regression trees (CART) into the framework of meta-analysis. This method identifies subgroups of homogeneous studies by searching influential moderators that can explain the heterogeneity, and performs subgroup analysis to test the significance of the identified moderators and estimate the subgroup effect sizes.

Methods: Meta-CART has the common limitations of tree-based methods: 1) the construction of a tree employs greedy search strategy, which does not guarantee a globally optimal solution and may lead to instable results; 2) the statistical inference in the identified subgroups is difficult, since the subgroups are not pre-defined but explored from the data. In this article, we propose extensions to overcome these limitations. For the tree construction, we propose using smooth sigmoid surrogate (SSS) and a look-ahead strategy to alleviate the local optimum problem and speed up the splitting procedure. For the statistical inference in the identified subgroups, we use permutation test to appropriately assess the statistical significance of the heterogeneity between identified subgroups, and propose a bootstrap procedure to correct the over-optimism in the confidence intervals of the estimated subgroup effect sizes. All these extensions are not restricted to meta-CART, and can be applied to any tree-based method that aims for subgroup identification.

Results: The simulation results show that applying both SSS method and the look-ahead

strategy improves the performance of meta-CART by reducing the false positive rate and improving the recovery rate. The permutation test improves the control of false positive findings with little sacrifice in recovery rates. The bootstrap procedure works well in correcting for the bias in the estimates of the residual heterogeneity.

Conclusion: This paper propose various extension for meta-CART to improve the stability of the tree construction process for subgroup identification, and to correct for potential bias in the statistical inference with respect to the identified subgroups. These extensions improve the ability of meta-CART to find the influential moderators and to identify the interaction effect among them, provide appropriate significance test for the moderator effects, and properly quantify the uncertainty in the estimates of effect sizes in the identified subgroups.

6.1 Introduction

In the last decades, the number of clinical trials has grown rapidly, and so did the need for systematic reviews and meta-analyses to summarize the research findings (Bastian, Glasziou, & Chalmers, 2010). Meta-analyses are typically used to assess the effectiveness of some intervention, which is done by computing a weighted average of the study effect sizes (i.e., an overall effect size). Furthermore, they can quantify the heterogeneity in the study effect sizes. For example, heterogeneity may be introduced by variation in the implementations of an intervention. Michie, Abraham, et al. (2009) showed in a meta-analysis that interventions designed to change health-related behaviors generally include various behavior change techniques that influence its effectiveness. Another possible source of heterogeneity arises when study samples are recruited from different populations. In a meta-analysis by Moyer, Rounds, and Hannum (2004) to examine the effectiveness of massage therapy, the samples of participants showed a large variety between studies such as pregnant women, surgery patients, healthy adults and so on. In these situations, it is of high interest to identify the study characteristics that explain the heterogeneity (i.e., moderators) and understand their influence on effect sizes. However, over the past decades, statistical approaches in meta-analysis primarily focused on confirmatory analyses, and method development aiming at explaining heterogeneity has been quite limited (Tipton et al., 2018). Conventionally, the mainly used method to examine the relationship between effect size and moderators is meta-regression. This method is based on linear modeling, and has difficulties in simultaneously investigating multiple moderators and examining interactions among them (Li, Dusseldorp, & Meulman, 2017). Therefore, as an alternative, some recent works have begun to use tree-based methods to investigate the source of heterogeneity (Bull et al., 2018; Dusseldorp et al., 2014; van Genugten et al., 2016).

In many fields, tree-based methods are popular approaches for subgroup identification purposes, because they excel at dealing with interactions and handling non-linear relationships. In addition, a fitted tree model mimics a hierarchically nested set of decision rules, and therefore can be easily interpreted. The most commonly used tree-based method is the classification and regression trees (CART) method proposed by Breiman et al. (1984), and there are many advanced variants of CART (for examples, see Dusseldorp et al., 2010; Dusseldorp & Van Mechelen, 2014; Foster, Taylor, & Ruberg, 2011; Lipkovich, Dmitrienko, Denne, & Enas, 2011; Loh, He, & Man, 2015; Su, Tsai, Wang, Nickerson, & Li, 2009). However, most of the methods are in single study settings, and the idea of applying tree-based methods in the framework of meta-analysis is relatively new. Dusseldorp et al. (2014) proposed a method called meta-CART that applies CART in meta-analysis for a health psychological study. This method was further developed by Li, Dusseldorp, and Meulman (2017); Li et al. (2019). Mistry et al. (2018) integrate the

subgroup identification based on different effect search (SIDES; Lipkovich et al., 2011) into the framework of individual patient data (IPD) meta-analysis. In the current paper we focus on the framework of aggregated meta-analysis and extend the meta-CART method.

The existing meta-CART method consists of three components: a procedure of splitting data (i.e., tree construction), a procedure of determining the best tree model (i.e., tree pruning), and a procedure of summarizing the terminal node (i.e., subgroup analysis). Some of the limitations of tree-based methods also apply to meta-CART, such as instability in the tree structure and difficulties in statistical inference for the subgroup analysis. Instability means that small variations in the data can result in very different trees. This problem is due to the greedy algorithm that is used and the hierarchical structure of trees. In the tree construction procedure, the greedy search algorithm optimizes the splitting criterion locally at each separate split. This process is "short-sighted" and may end up in a local optimum. If an error occurs in the top split, its influence will be propagated down to all the splits below because of the hierarchical structure, and this constitutes an inherent instability (see Friedman et al., 2001, p. 274). To reduce this variability, ensemble learning methods such as Bagging (Breiman, 1996), Boosting (Friedman, 2001), and Random Forests (Breiman, 2001) are often used in the field of data mining by combining or averaging trees built, for example, on re-sampled data. However, the resulting model, which is an ensemble of multiple trees, loses the interpretability of a single tree, and the influence of the predictor variables becomes unclear. Besides, the sample size (i.e., the number of studies) in meta-analysis is usually not large enough for ensemble learning. Thus, ensemble learning methods do not serve well the purpose of meta-CART analysis to understand the influence of moderators. The difficulties in statistical inference are due to the fact that the identified subgroups are not pre-defined but explored from the data. In the subgroup analysis procedure, the subgroup effect sizes are estimated and the significance of between-subgroups heterogeneity (i.e., whether the identified subgroup memberships account for a significant proportion of the variance in study effect sizes) is tested. Since the subgroup memberships are identified and tested using the same data, it raises post hoc issues and problems of statistical inference. Thus, the heterogeneity test cannot be interpreted as its face value and only gives information when the test results are not significant (see Li et al., 2019, Section 8.1). Also, due to the adaptive nature of tree modeling, over-optimism exists in the confidence intervals of the subgroup effect sizes, and the confidence intervals may be too narrow to have the desired coverage.

These limitations motivate us to extend the existing meta-CART method in four aspects: two of these are in the tree construction stage to increase stability and computational efficiency, and two are in the subgroup analysis stage to correct for over-optimism. These four extensions can also be applied to other tree-based methods aiming at subgroup identification (e.g., Lipkovich, Dmitrienko, & B D'Agostino Sr, 2017). First, we

use the smooth sigmoid surrogate (SSS) strategy (Su et al., 2016; Su, Peña, Liu, & Levine, 2018) as an alternative for the greedy search. Second, we propose a novel approach that combines SSS with a look-ahead strategy. Third, we use a permutation test as a more appropriate test to assess whether the identified moderators are spurious. Fourth, we propose a new bootstrap-based approach to estimate and correct the bias in the confidence intervals of subgroup effect sizes.

The outline of this paper is as follows. First, we describe shortly the existing meta-CART method. Second, we introduce the four new approaches to extend meta-CART. Third, we evaluate the performance of the new approaches in a simulation study. Next we illustrate the new extended meta-CART with a real-world data set. Finally, we summarize and discuss the results.

6.2 The Existing meta-CART Method

In this section, we briefly describe the three components of meta-CART followed by a description of the splitting criterion.

In a meta-CART analysis, the data consist of observations from K studies $\mathcal{D} = \{(g_k, \sigma^2_{\epsilon_k}, x_k : k = 1, \ldots, K)\}$, where g_k and $\sigma^2_{\epsilon_k}$ are the estimated effect size and the sampling variance in the k^{th} study, and $x_k = (x_{k1}, \ldots, x_{kM})^T \in \mathbb{R}^M$ is a M-dimensional covariate vector corresponding to the characteristics of the k^{th} study (i.e., potential moderators). There are two general approaches to a meta-CART analysis: fixed effect (FE) and random effects (FE) assumptions (Li et al., 2019). The main difference between these two assumptions is that FE meta-CART ignores the residual heterogeneity unexplained by the moderators and weights the studies by the inverse of the sampling variance ($w_k = 1/\sigma^2_{\epsilon_k}$), while RE meta-CART takes into account the residual heterogeneity σ^2_τ and defines weights as $w^*_k = 1/(\sigma^2_{\epsilon_k} + \sigma^2_\tau)$.

In the tree construction procedure, meta-CART follows the paradigm of CART and splits the data in such a way that the heterogeneity between offspring nodes is maximized or, equivalently, the within-node heterogeneity is minimized. At each split, a parent node is partitioned into two child nodes by applying a threshold on a selected moderator X_m at the corresponding split point c. When X_m is an ordinal variable, a split of the data is introduced by asking a binary question such as "Is X_m greater than c?". When X_m is a categorical variable, the binary question becomes " Does X_m belongs to $\mathcal{C}$?", where $\mathcal{C}$ is a subset of the variable levels. For computational efficiency, one common strategy is to treat categorical levels as an ordinal variable by sorting the variable levels according to the group means (see Breiman et al., 1984, Sections 4.2 and 9.4). To find X_m and c, the greedy search compares all possible splits with respect to the splitting criterion and selects the split that optimizes this criterion. The splitting algorithm is fully recursive in FE meta-CART. But RE meta-CART applies a sequential splitting algorithm that

refreshes the global estimate of σ_τ^2 after each split. The "global" estimate means that σ_τ^2 is estimated across all terminal nodes rather than only within the resulting left and right child nodes after a split. For both assumptions, the splitting process continues until all terminal nodes contain only a few studies, and a large initial tree is constructed as a result.

In the tree pruning procedure, the initial tree is pruned back to a smaller size to prevent overfitting. This procedure consists of three steps. First, the initial tree is iteratively pruned to a nested sequence of subtrees. Second, ten-fold cross-validation is performed to estimate the prediction errors for each subtree. Then the final tree is selected as the smallest tree with a cross-validation error that is within the minimum cross-validation error plus half the standard error. [1]

The terminal nodes of the final tree form the subgroups of the studies. In the final step, a standard subgroup analysis is performed to assess the relationship between the identified subgroup membership and the effect size. The subgroup analysis results consist of two parts: an overall heterogeneity test for the identified subgroups, and the summary effect size in each subgroup. The significance test is based on the between-subgroups heterogeneity, which is also called the between-subgroups Q-statistic. The computation of this Q-statistic will be described in Section 6.2.1. The subgroup effect sizes are estimated as weighted means. Denote the final tree by $\mathcal{T}$, and let $\widetilde{\mathcal{T}}$ be the set of all the terminal nodes in the final tree $\mathcal{T}$. For FE meta-CART, the within-node summary effect size is computed as

$$g_{t+} = \frac{\sum_{k\in t} w_k g_k}{\sum_{k\in t} w_k} \tag{6.1}$$

for each terminal node $t \in \widetilde{\mathcal{T}}$.

The $(1-\alpha)\times 100\%$ confidence interval (CI) is constructed as:

$$g_{t+} \pm z_{1-\alpha/2}\, \sigma_t, \tag{6.2}$$

where $z_{1-\alpha/2}$ is the $(1-\alpha/2)$-th percentile of the standard normal $\mathcal{N}(0,1)$ distribution and σ_t denotes the standard error (SE) of the estimated effect size in node t. SE is computed as $\sigma_t = 1/\sqrt{\sum_{k\in t} \sigma_{\epsilon_k}^2}$.

For RE meta-CART,

$$g_{t+}^* = \frac{\sum_{k\in t} w_k^* g_k}{\sum_{k\in t} w_k^*}. \tag{6.3}$$

The $(1-\alpha)\times 100\%$ confidence interval (CI) is constructed as:

$$g_{t+}^* \pm z_{1-\alpha/2}\, \sigma_t^*, \tag{6.4}$$

[1] This half-standard-error rule is recommended by Li et al. (2019). Due to the relatively small sample sizes (i.e., number of studies) in a meta analysis, the half-standard-error rule performs better than the one-standard-error rule in meta-CART.

where

$$\sigma_t^* = 1/\sqrt{\sum_{k \in t} (\sigma_{\epsilon_k}^2 + \sigma_\tau^2)}. \tag{6.5}$$

6.2.1 The Splitting Criterion

The splitting criterion differs for the FE and RE meta-CART algorithms. We start with FE meta-CART, which maximizes the Q-statistic between the left and right child nodes after each split. The Q-statistic is defined as

$$Q_B = \sum_{k \in t_L} w_k (g_{t_L+} - g_{t+})^2 + \sum_{k \in t_R} w_k (g_{t_R+} - g_{t+})^2, \tag{6.6}$$

where g_{t+} is the weighted mean of the parent node t with weights $w_k = 1/\sigma_{\epsilon_k}^2$, and t_L and t_R are the resulting left and right child nodes.

RE meta-CART estimates σ_τ^2 and selects the split that maximizes the Q-statistic between all terminal nodes after each split. While several methods are available, we choose to compute the residual heterogeneity σ_τ^2 by using the DerSimonian and Laird method (DerSimonian & Laird, 1986):

$$\sigma_\tau^2 = \frac{\sum_{t \in \mathcal{T}} \sum_{k \in t} w_k (g_k - g_{t+})^2 - \sum_{t \in \mathcal{T}} (K_t - 1)}{\sum_{t \in \mathcal{T}} C_t}, \tag{6.7}$$

where $\mathcal{T}$ is the entire tree after the split, K_t is the number of studies in node t, and the components C_t are computed as

$$C_t = \sum_{k \in t} w_k - \frac{\sum_{k \in t} w_k^2}{\sum_{k \in t} w_k}. \tag{6.8}$$

Given the estimated σ_τ^2, the between-subgroups Q-statistic is computed as

$$Q_B^* = \sum_{t \in \mathcal{T}} \sum_{k \in t} w_k^* (g_{t+}^* - g_{++}^*)^2, \tag{6.9}$$

where g_{++}^* is the grand weighted mean with weights $w_k^* = 1/(\sigma_{\epsilon_k}^2 + \sigma_\tau^2)$.

6.3 New Extensions for Meta-CART

6.3.1 The Smooth Sigmoid Surrogate to Identify the Best Split Point

Since a split is made on a selected moderator X_m at its split point c, we can introduce a binary variable $\delta_k = \delta(x_{km}; c) = I(x_{km} > c)$, and write the summary effect size in the

two resulting child nodes as

$$\begin{cases} g_{t_L+} = \frac{\sum_{k=1}^{K_t}(1-\delta_k)w_k g_k}{\sum_{k=1}^{K_t}(1-\delta_k)w_k}, \\ g_{t_R+} = \frac{\sum_{k=1}^{K_t}\delta_k w_k g_k}{\sum_{k=1}^{K_t}\delta_k w_k}. \end{cases} \tag{6.10}$$

The FE Q-statistic in (6.6) can be re-written as

$$Q_B = (g_{t_L+} - g_{t+})^2 \sum_{k=1}^{K_t}(1-\delta_k)w_k + (g_{t_R+} - g_{t+})^2 \sum_{k=1}^{K_t}\delta_k w_k. \tag{6.11}$$

To find the best split point $\hat{c}$ for X_m, a greedy search computes Q_B at all possible values of c and selects the one resulting in the largest increase of Q_B. This discrete process tends to produce large fluctuations with local spikes (Su et al., 2018). Also, a greedy search can be time-consuming when X_m has many distinct values. To overcome these difficulties, we use the smooth sigmoid surrogate (SSS) strategy (Su et al., 2016, 2018) as an approximation method for a greedy search. There are many choices of the sigmoid function to approximate the threshold indicator function δ_k. Here we use the logistic function

$$s(x_{km}; a, c) = \frac{\exp\{a(x_{km} - c)\}}{1 + \exp\{a(x_{km} - c)\}}, \tag{6.12}$$

with scale parameter $a > 0$. Then the summary effect sizes d_{t_L+} and d_{t_L+} can be approximated by

$$\begin{cases} \tilde{d}_{t_L+} = \frac{\sum_{k=1}^{K_t}(1-s_k)w_k d_k}{\sum_{k=1}^{K_t}(1-s_k)w_k}, \\ \tilde{d}_{t_R+} = \frac{\sum_{k=1}^{K_t}s_k w_k d_k}{\sum_{k=1}^{K_t}s_k w_k}. \end{cases} \tag{6.13}$$

We approximate the splitting criterion Q_B by

$$\tilde{Q}_B = (\tilde{g}_{t_L+} - g_{t+})^2 \sum_{k=1}^{K_t}(1-s_k)w_k + (\tilde{g}_{t_R+} - g_{t+})^2 \sum_{k=1}^{K_t}s_k w_k. \tag{6.14}$$

The value of a can be fixed a priori, so the best split point $\hat{c}$ is obtained by maximizing $\tilde{Q}_B$, which is a smooth object function for c only. Su et al. (2016, 2018) recommend to fix a at a value in the range $[10, 50]$. In order to do so, moderator X_m needs to be standardized. After solving the optimization problem, the identified $\hat{c}$ can be transformed back to the original scale for interpretability. We solve the one-dimensional smooth optimization problem by using the Brent method (Brent, 2013), which is implemented in R in the function `optimize()`. Options for multiple starts and constraining the search range are available to come close to the global optimum and avoid the "end-cut preference" problem (see Breiman et al., 1984, Section 11.8).

The SSS strategy is applied in RE meta-CART using the same idea as described above.

The key difference is that the estimation of σ_τ^2 is also a function of the threshold indicator function δ_k.

6.3.2 The Look Ahead Strategy

Look ahead with a greedy search

The tree construction process has a greedy nature: it chooses each split with no regard of future splits. Thus, it does not guarantee a globally optimal solution for the entire tree. To alleviate this problem, we propose a look-ahead strategy that looks one step further when choosing the splits.

Starting the algorithm at the root node, a look-ahead strategy searches for the optimal combination of a split of the root node and a split of one of its child nodes. The splitting criterion is evaluated at all possible combinations of two split points (i.e., single split points on any two moderators or any two split points on one moderator). The first split is chosen based on the combination that maximizes the splitting criterion. Then the look-ahead strategy can be repeated for the resulting child nodes. Because the computational burden grows exponentially as the number of splits to be examined increases (see Esmeir & Markovitch, 2007), we recommend applying this look-ahead procedure only for a few initial splits (i.e., no more than five splits). After the look-ahead procedure, the resulting offspring nodes can be further split following a fully greedy procedure for maximizing the splitting criterion at each split.

Look ahead with SSS

As an alternative for a greedy search, SSS can be applied in the look-ahead strategy to reduce the variability introduced by local spikes and the computational burden. Instead of looking at all combinations of possible split points, the look-ahead with SSS strategy estimates the optimal split points for each combination of moderators, and selects the combination that results in the largest value of the splitting criterion.

When making two splits from the root node, there are two possible scenarios: (a) the second split is from the left child node (Figure 6.1a), and (b) the second split is from the right child node (Figure 6.1b). Denote the three grandchild nodes by $t_{G_1}, t_{G_2}, t_{G_3}$. Let X_1 be the first splitting moderator and c_1 its split point, and X_2 and c_2 be the second splitting moderator and its split point. For scenario (a), the subgroup effect sizes are

computed as

$$\begin{cases} g_{t_{G_1}+} = \dfrac{\sum_{k=1}^{K} I(x_{k1} \le c_1; x_{k2} \le c_2) \cdot w_k g_k}{\sum_{k=1}^{K} I(x_{k1} \le c_1; x_{k2} \le c_2) \cdot w_k} & \approx \tilde{g}_{t_{G_1}+} = \dfrac{\sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\}\{1 - s(x_{k2}; a, c_2)\} \cdot w_k g}{\sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\}\{1 - s(x_{k2}; a, c_2)\} \cdot w_k} \\ g_{t_{G_2}+} = \dfrac{\sum_{k=1}^{K} I(x_{k1} \le c_1; x_{k2} > c_2) \cdot w_k g_k}{\sum_{k=1}^{K} I(x_{k1} \le c_1; x_{k2} > c_2) \cdot w_k} & \approx \tilde{g}_{t_{G_2}+} = \dfrac{\sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\} s(x_{k2}; a, c_2) \cdot w_k g_k}{\sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\} s(x_{k2}; a, c_2) \cdot w_k}, \\ g_{t_{G_3}+} = \dfrac{\sum_{k=1}^{K} I(x_{k1} > c_1) \cdot w_k g_k}{\sum_{k=1}^{K} I(x_{k1} > c_1) \cdot w_k} & \approx \tilde{g}_{t_{G_3}+} = \dfrac{\sum_{k=1}^{K} s(x_{k1}; a, c_1) \cdot w_k g_k}{\sum_{k=1}^{K} s(x_{k1}; a, c_1) \cdot w_k}. \end{cases} \tag{6.15}$$

The splitting criterion is approximated by

$$\begin{aligned} \tilde{Q}_B &= (\tilde{g}_{t_{G_1}+} - g_{++})^2 \sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\}\{1 - s(x_{k2}; a, c_2)\} \cdot w_k \\ &\quad + (\tilde{g}_{t_{G_2}+} - g_{++})^2 \sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\} s(x_{k2}; a, c_2) \cdot w_k. \\ &\quad + (\tilde{g}_{t_{G_3}+} - g_{++})^2 \sum_{k=1}^{K} s(x_{k1}; a, c_1) \cdot w_k. \end{aligned} \tag{6.16}$$

For scenario (b), the subgroup effect sizes can be computed by

$$\begin{cases} g_{t_{G_1}+} = \dfrac{\sum_{k=1}^{K} I(x_{k1} \le c_1) \cdot w_k g_k}{\sum_{k=1}^{K} I(x_{k1} \le c_1) \cdot w_k} & \approx \tilde{g}_{t_{G_1}+} = \dfrac{\sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\} \cdot w_k g_k}{\sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\} \cdot w_k}, \\ g_{t_{G_2}+} = \dfrac{\sum_{k=1}^{K} I(x_{k1} > c_1; x_{k2} \le c_2) \cdot w_k g_k}{\sum_{k=1}^{K} I(x_{k1} > c_1; x_{k2} \le c_2) \cdot w_k} & \approx \tilde{g}_{t_{G_2}+} = \dfrac{\sum_{k=1}^{K} s(x_{k1}; a, c_1)\{1 - s(x_{k2}; a, c_2)\} \cdot w_k g_k}{\sum_{k=1}^{K} s(x_{k1}; a, c_1)\{1 - s(x_{k2}; a, c_2)\} \cdot w_k}, \\ g_{t_{G_3}+} = \dfrac{\sum_{k=1}^{K} I(x_{k1} > c_1; x_{k2} > c_2) \cdot w_k g_k}{\sum_{k=1}^{K} I(x_{k1} > c_1; x_{k2} > c_2) \cdot w_k} & \approx \tilde{g}_{t_{G_3}+} = \dfrac{\sum_{k=1}^{K} s(x_{k1}; a, c_1) s(x_{k2}; a, c_2) \cdot w_k g_k}{\sum_{k=1}^{K} s(x_{k1}; a, c_1) s(x_{k2}; a, c_2) \cdot w_k}. \end{cases} \tag{6.17}$$

The splitting criterion is approximated by

$$\begin{aligned} \tilde{Q}_B &= (\tilde{g}_{t_{G_1}+} - g_{++})^2 \sum_{k=1}^{K} \{1 - s(x_{k1}; a, c_1)\} \cdot w_k \\ &\quad + (\tilde{g}_{t_{G_2}+} - g_{++})^2 \sum_{k=1}^{K} s(x_{k1}; a, c_1)\{1 - s(x_{k2}; a, c_2)\} \cdot w_k. \\ &\quad + (\tilde{g}_{t_{G_3}+} - g_{++})^2 \sum_{k=1}^{K} s(x_{k1}; a, c_1) s(x_{k2}; a, c_2) \cdot w_k. \end{aligned} \tag{6.18}$$

For each scenario, $\hat{c_1}$ and $\hat{c_2}$ are obtained by maximizing $\tilde{Q}_B$. We solve this two-dimensional optimization problem by using the limited-memory Broyden-Fletcher-Goldfarb-Shanno box-constraints (L-BFGS-B) method (Byrd, Lu, Nocedal, & Zhu, 1995). The L-BFGS-B method belongs to the quasi-Newton methods, and can solve large nonlinear optimization

problems with simple box constraints. The method is implemented in R by the function `optim()`, and options for multiple starts and constraining the search range are available.

After solving for the best split points in both scenarios, the look-ahead with SSS strategy chooses the one that results in the higher value for $\tilde{Q}_B$. Summarizing, a look-ahead strategy with SSS consists of three sub-procedures. The first finds all possible combinations of two moderators [2]. Within each combination, the second sub-procedure finds the best split points according to the two splitting scenarios, and selects the scenario with higher values for the between-subgroups Q-statistic. Finally, the third sub-procedure compares the selected scenario and split points across all combinations, and selects the combination that results in the highest value for the between-subgroups Q-statistic.

6.3.3 A Permutation Test for Between-Subgroups Heterogeneity

After the tree model is pruned via cross-validation, a subgroup meta-analysis is performed to estimate the subgroup effect sizes and to test the significance of the between subgroups heterogeneity (i.e., whether the identified tree model accounts for a significant proportion of the variance in study effect sizes). In a standard meta-analysis, the heterogeneity test, often called Q-test, is based on the between-subgroups Q-statistic, which asymptotically follows a chi-square distribution under the null hypothesis. However, this standard Q-test is inappropriate in meta-CART analysis, since the subgroup membership results from the tree model that maximizes the same Q-statistic. To solve this problem, we propose to use a permutation test to correct for the over-optimism in the Q-test. This provides a stricter way to determine whether moderators effects, represented by the tree model, are spurious. This permutation test works as is shown in Algorithm 1.

6.3.4 Bias Correction via Bootstrapping

As addressed in the previous section, the subgroups obtained by a meta-CART analysis are not pre-defined but found by the tree algorithm. Thus, the within-subgroup heterogeneity is underestimated, and over-optimism exists in the standard errors of the subgroup effect sizes. To appropriately quantify the uncertainty in the estimates for the subgroup effect sizes, the bias in the within-node heterogeneity needs to be estimated. Note that this is only necessary for RE meta-CART, since there is no within-node heterogeneity under

[2]Note that two splits on the same moderator are allowed.

[3]For FE meta-CART, if the optimal subtree with G terminal nodes is not available by pruning, we select a smaller subtree that has the closest number of terminal nodes, and then further split the selected subtree until the number of terminal nodes reaches G. For RE meta-CART, because of its non-recursive nature, any pruning will change the estimate of residual heterogeneity and the Q-statistic for each subtree. Therefore, we sequentially grow the tree until the number of terminal nodes reaches G, and no pruning procedure is applied.

Data: Tree with terminal nodes $t_1, t_2, ..., t_G$ constructed from
$\mathcal{D} = \{d_k, \sigma^2_{\epsilon_k}, \mathbf{X}_k, k = 1, 2, ..., K\}$.
Result: p-value for the Q-statistic between the terminal nodes.
begin
  compute Q between $t_1, t_2, ..., t_G$;
  $\gamma \leftarrow 0$;
  for $b \leftarrow 1$ *to* B **do**
    permute $\mathcal{D}^*_p = \{d^*_k, \sigma^*_{\epsilon_k}, \mathbf{X}_k, k = 1, 2, ..., K\}$ from $\mathcal{D}$ with d_k and $\sigma^2_{\epsilon_k}$ bounded together and their correspondence with $\mathbf{X}_k$ under random permutation;
    construct from $\mathcal{D}^*_p$ the optimal subtree with terminal nodes $t^*_1, t^*_2, ..., t^*_G$[3];
    compute Q^*_p between $t^*_1, t^*_2, ..., t^*_G$;
    if $Q^*_p \geq Q$ **then**
      $\gamma \leftarrow \gamma + 1$;
    else
      $\gamma \leftarrow \gamma$;
    end
  end
  $p = (\gamma + 1)/(B + 1)$
end

Algorithm 1: Permutation significance test for between-subgroups heterogeneity.

the FE assumption; in other words, the optimism in the SE estimates of FE meta-CART is due to the FE assumption. One common method for bias correction is the bootstrap (Efron & Tibshirani, 1994). In this approach, the estimates from the bootstrap samples are averaged and are compared to the original estimate to give an estimator of the bias. However, there is one major obstacle to apply the bootstrap in tree-based methods. Trees are unstable in the sense that a small perturbation of the data often results in a substantially different tree model. As a result, the tree models obtained by using bootstrap samples are different from each other and from the final tree model constructed from the original sample. To tackle this problem, we note that a tree model forms a natural grouping of the entire data. With two tree structures, observations in a node from one tree can be distributed into different nodes of the other tree. Utilizing this property, we propose an appropriate bootstrap bias correction procedure for the underestimated within-node heterogeneity as outlined in Algorithm 2.

input : data $\mathcal{D} = \{d_k, \sigma_{\epsilon_k}, \mathbf{X}_k, k = 1, 2, ..., K\}$.
output: A tree model $\mathcal{T}$ with bias corrected within-node heterogeneity Q_t for each $t \in \widetilde{\mathcal{T}}$.
initialize $B = \#$ bootstrap samples;
begin
construct a best-sized tree $\mathcal{T}$ from $\mathcal{D}_b$ via pruning and cross validation;
obtain node membership vector $\mathbf{m}_0 \in \mathbb{R}^K$ for all studies in $\mathcal{D}$ w.r.t. $\mathcal{T}$;
compute μ_t for each $t \in \widetilde{\mathcal{T}}$ based on $\mathcal{D}$;
set bias $b_t = 0$ for $t \in \widetilde{\mathcal{T}}$;
for $b \leftarrow 1$ **to** B **do**
draw a bootstrap sample $\mathcal{D}_b$;
construct a best-sized tree $\mathcal{T}_b$ from $\mathcal{D}_b$ via pruning and cross validation;
compute heterogeneity averaged over weights
$\{\mu_{bt'} = \sum_{k \in t'} w_k^* (d_k - d_{t'+})^2 / \sum_{k \in t'} w_k^* : t' \in \widetilde{\mathcal{T}}_b\}$ based on $\mathcal{D}_b$;
send $\mathcal{D}$ down to $\mathcal{T}_b$ and recompute compute weighted average of squares
$\{\mu_{0t'} : t' \in \widetilde{\mathcal{T}}_b\}$ on basis of $\mathcal{D}$;
compute bias $b_{bt'} = \mu_{0t'} - \mu_{bt'}$ for $t' \in \widetilde{\mathcal{T}}_b$;
`/* see how studies in` $t \in \widetilde{\mathcal{T}}_0$ `are distributed over` $\widetilde{\mathcal{T}}_b$. `*/`
obtain node membership vector $\mathbf{m}_b \in \mathbb{R}^n$ for all studies in $\mathcal{D}$ w.r.t. $\mathcal{T}_b$;
form two-way contingency table of weights
$\{m_{tt'}^* = \sum_{k \in t \text{ and } k \in t'} w_k^* : t \in \widetilde{\mathcal{T}}$ and $t' \in \widetilde{\mathcal{T}}_b\}$ with $\mathbf{m}_0$ and $\mathbf{m}_b$;
compute weighted row proportions $p_{tt'}^* = m_{tt'}^* / m_{t.}^*$;
for $t \in \widetilde{\mathcal{T}}$ **do**
update $b_t := b_t + \sum_{t' \in \widetilde{\mathcal{T}}_b} p_{tt'}^* b_{bt'}$;
end
end
average bias $b_t := b_t / B$ for $t \in \widetilde{\mathcal{T}}$;
bias correction $Q_t := (\mu_t + b_t) \cdot \sum_{k \in t} w_k^*$ for $t \in \widetilde{\mathcal{T}}$.
end

Algorithm 2: Bias correction for within-node heterogeneity in tree modeling.

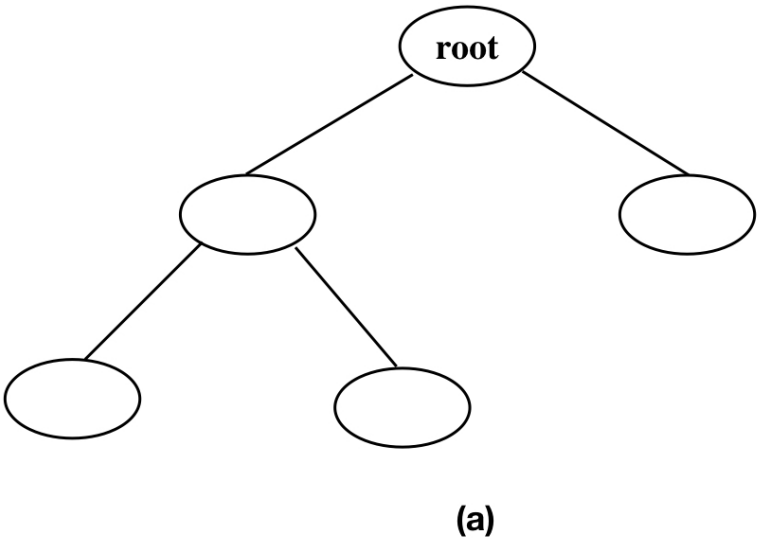
Specifically, let $\mathbf{m}_0 \in \mathbb{R}^K$ denote the node membership vector that assigns a terminal node of $\mathcal{T}$ to each observation in $\mathcal{D}$. This node membership vector $\mathbf{m}_0$ is categorical with levels $\{1, 2, \ldots, |\widetilde{\mathcal{T}}|\}$. We take B bootstrap samples $\{\mathcal{D}_b : b = 1, \ldots, B\}$. For each bootstrap sample $\mathcal{D}_b$, a best-sized tree $\mathcal{T}_b$ is constructed and we obtain the average weighted sum of squares $\{\mu_{bt'} : t' \in \widetilde{\mathcal{T}}_b\}$. Sending $\mathcal{D}$ down to $\mathcal{T}_b$ and recomputing $\{\mu_{0t'} : t' \in \widetilde{\mathcal{T}}_b\}$, based on $\mathcal{D}$, yields estimates of the bias $\{b_{bt'} = \mu_{0t'} - \mu_{bt'} : \ t' \in \widetilde{\mathcal{T}}_b\}$.

To obtain bias estimates b_t for each μ_t in $\{\mu_t : t \in \widetilde{\mathcal{T}}\}$, $b_{bt'}$ is weighted by looking at how observations in $t \in \widetilde{\mathcal{T}}$ are distributed over $\widetilde{\mathcal{T}}_b$. To proceed, let $\mathbf{m}_b \in \mathbb{R}^K$ denote

the node membership vector that assigns a terminal node of $\mathcal{T}_b$ to each observation in $\mathcal{D}$. The two categorical vectors $\mathbf{m}_0$ and $\mathbf{m}_b$ form a $|\widetilde{\mathcal{T}}| \times |\widetilde{\mathcal{T}}_b|$ two-way contingency table of weights $\{m^*_{tt'} = \sum_{k \in t \text{ and } k \in t'} w^*_k : t \in \widetilde{\mathcal{T}} \text{ and } t' \in \widetilde{\mathcal{T}}_b\}$. Weights are applied here to take the accuracy of the estimated effect size in each study into consideration. Let $p^*_{tt'} = m^*_{tt'}/m^*_{t\cdot}$ be the weighted row marginal proportions, where $m^*_{t\cdot} = \sum_{t'} m^*_{tt'}$ is the t-th row total. Then an estimate of the bias from the bth bootstrap sample $\mathcal{D}_b$ is given by $\sum_{t' \in \widetilde{\mathcal{T}}_b} p^*_{tt'} b_{bt'}$. Averaging over B bootstrap samples leads to the bias estimate for μ_t and bias correction on μ_t can be made accordingly. Put together, the bias-corrected within-node heterogeneity Q_t is given by

$$Q_t := \{\mu_t - \frac{1}{B} \sum_{i=1}^{B} \sum_{t' \in \widetilde{\mathcal{T}}_b} p^*_{tt'} (\mu_{0t'} - \mu_{bt'})\} \cdot \sum_{k \in t} w^*_k. \tag{6.19}$$

With the bias-corrected Q_t, the residual heterogeneity σ^2_τ is re-estimated using (6.7). And the SEs of subgroup effect sizes are corrected using (6.5).



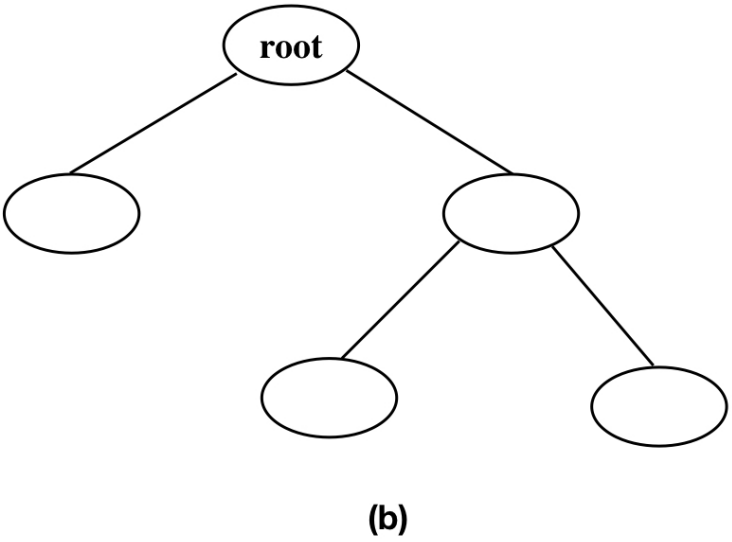


Figure 6.1: Two possible scenarios to grow a tree with two splits from the root node.

6.4 Design of the Simulation Studies

In the following simulation studies, we aim at evaluating the performance of the proposed extensions for meta-CART. First, we compare the ability to retrieve the true model underlying the data for four implementations of meta-CART: 1) the standard FE meta-CART, 2) FE meta-CART applying SSS and the look-ahead strategy, 3) the standard RE meta-CART, and 4) RE meta-CART applying SSS and the look-ahead strategy. Second, we apply the permutation test in the four implementations of meta-CART and evaluate its influence on the recovery performance. Third, we apply the bootstrap correction procedure and evaluate whether the over-optimism is properly reduced.

The simulation studies are performed in the R language (Team, 2017). The meta-CART analyses are performed using our R-package `metacart3`, which is available at `https://github.com/XinruLI/metacartv3`.

6.4.1 Data Generation

In the previous studies by Li, Dusseldorp, and Meulman (2017); Li et al. (2019), it was shown that the performance of meta-CART can be influenced by many design factors, such as the number of studies K, the within-study sample size $\bar{n}$, the number of potential moderators M, the magnitude of the interaction effect size g_I, the common correlation between the moderators ρ, the residual heterogeneity σ_τ^2, and the complexity of the true model underlying the data. In this study, we focus on three design factors: the number of studies $K = \{40, 120\}$, the residual heterogeneity $\sigma_\tau^2 = \{0, 0.025, 0.05\}$, and the five models (A-E) to generate the data differing in complexity as shown in Table 6.1. We include the number of studies because it is the most interested factor in practice. The residual heterogeneity is included because this factor is essential to compare the FE and RE models. The complexity of the true model to generate the data is included because it has the largest influence on the ability of meta-CART to retrieve the true structure underlying the data. The other design factors are fixed at the medium level: $\bar{n} = 80$, $M = 10$, $g_I = 0.5$, $\rho = 0.3$. For each cell of the design, 1000 data sets were generated as training sets. As a result, $2 \times 3 \times 5 \times 1000 = 30,000$ training sets were generated. In addition, to estimate the prediction performance and the over-optimism in the subgroup analysis results, a test set with $K = 2000$ studies is generated for each cell of the design.

For each data set, 10 moderators are generated from the uniform distribution ($x_m \sim$ uniform[0,1] for $m = 1, \ldots, 10$) with a common correlation $\rho = 0.3$. To achieve this, first multivariate normal vectors with the common correlation ρ are simulated, and then the probability integral transform is applied to the simulated vectors (Gilli, Maringer, & Schumann, 2011).

Five models are used to generate true effect sizes given a selection of moderators. The models vary in the number of moderators involved. Note that we generated ten

Table 6.1: The five models used to generate data sets.

Model	Form
A	$g(x) = 0.5$
B	$g(x) = 0.5 \cdot I(x_1 > 0.5)I(x_2 > 0.5)$
C	$g(x) = 0.5 \cdot I(x_1 > 0.5)I(x_2 \leq 0.5) + 0.5 \cdot I(x_1 \leq 0.5)I(x_2 > 0.5)$
D	$g(x) = -0.5 + 0.5x_1 + 0.5x_2$
E	$g(x) = -1.5 + sin(\pi x_1 x_2) + 2(x_3 - 0.5)^2 + x_4 + 0.5x_5$

moderators in each data set. Therefore, the moderators that are not involved in the true model are just noise factors. Model A is a null model to assess the false positive rate (meta-CART falsely identifies moderator effects when there is no moderator in the true model). Models B and C are used to evaluate the ability of meta-CART to identify the interaction effect(s) when the underlying structure can be fully expressed by a binary tree. Model B is a two-way interaction effect between two moderators. Model C includes two balanced two-way interactions, which is similar to the simulated example used by Tibshirani and Knight (1999). This model is shown by these authors to be difficult for greedy search procedures like CART because there is no information on where to split at the top level. Thus, the standard meta-CART is likely to suffer a local optimum problem in this case. Models D and E are used to evaluate the ability of meta-CART to identify influential moderators when the underlying structure cannot be fully captured by a binary tree. Model D is a linear model with main effects of two moderators. Model E is a nonlinear model derived from Friedman (1991). The coefficients in model E are rescaled so that the resulting effect size ranges between -1.2 and 1.2, which is commonly encountered in practice.

For a single study, the effect size was sampled from a normal distribution with mean $\bar{g}(x_k)$, computed as in Table 6.1 and variance σ_τ^2. We use the same method as in Viechtbauer (2007b) to generate the within-study sample size n_k from a normal distribution with mean $\bar{n}$ and standard deviation $\bar{n}/3$. Finally, given the values of n_k and $\bar{g}(x_k)$, the observed effect size g_k was sampled from a non-central t-distribution (for details, see Appendix A.1).

6.4.2 Evaluation of recovery performance

The four implementations of meta-CART are applied to the data sets that were generated. Once a non-trivial tee is identified, a permutation test with 500 permuted samples ($P = 500$) is performed to evaluate the significance of the heterogeneity between the identified subgroups (i.e., the Q-between). Due to the computational cost, the look-ahead strategy is only applied to the root node (i.e., for the first two splits). For all implementations of meta-CART, the half-standard-error rule is applied in the pruning procedure. The

following criteria are used to evaluate the recovery performance:

Criterion 1. Meta-CART falsely detects the presence of (a) moderator effect(s) in the data sets generated from model A (false positive rate). This criterion is computed as the percentage of meta-CART solutions within a cell where a non-trivial tree is found and if the heterogeneity test is significant.

Criterion 2. Meta-CART successfully selects all moderators used in the true model (full recovery rate). This criterion examines if the true model is fully recovered with all the true moderators and if no spurious moderators are selected. It is computed as the percentage of the true moderators being correctly identified by meta-CART to have significant heterogeneity test results.

Criterion 3. Meta-CART selects part of the moderators used in the true model (partial recovery rate), and does not select any spurious moderators. Criterion 3 is computed as the percentage of the true moderators being partially identified by meta-CART to have significant heterogeneity test results.

Criterion 4. The trained meta-CART models are applied to predict the study effect sizes for the test set, and the prediction errors, measured by mean squared error (MSE), are computed.

6.4.3 Evaluation of the bootstrap correction

Due to the computational cost, we use a part of the generated data sets to investigate the performance of the bootstrap correction. Among the training sets that resulted in non-trivial trees, we randomly select 100 non-trivial solutions to estimate the residual heterogeneity in both training and the corresponding test sets that are generated from the same underlying structure. Then we applied Algorithm 2 to each selected training set $\mathcal{D}$, and the resulting tree model $\mathcal{T}$, with the number of bootstrap samples $B = 50$, as LeBlanc and Crowley (1993) suggest using $25 \leq B \leq 100$. As a result, for each of the four selected cells, we obtain 100 estimates of residual heterogeneity in the training sets, 100 estimates of residual heterogeneity in the test sets, and 100 estimates of residual heterogeneity after bootstrap correction in the training sets. If the residual heterogeneity is properly estimated, the latter estimates in the training sets should be close to their corresponding estimates in the test sets.

6.5 Results of the Simulation Studies

6.5.1 SSS and Look-ahead strategy

The false positive rates of the four implementations of meta-CART can be found in Table 6.2. It is shown that the false positive rates are mainly influenced by the number of studies K, and the various implementations of meta-CART. The false positive rates decrease with the increase of K. Before correction by the permutation test, the two implementations of RE meta-CART have larger false positive rates than FE meta-CART. After the correction, the difference is small. For both FE and RE meta-CART, applying with SSS and look-ahead strategy reduces the rates of identifying non-trivial trees when there is no moderator effect in the true model (model A).

Figure 6.2 presents the full recovery rates and the partial recovery rates of the four implementations of meta-CART. In general, RE meta-CART applying SSS and the look-ahead strategy results in the highest recovery rates in most scenarios, and it outperforms FE meta-CART even when $\sigma_\tau^2 = 0$. In particular, applying SSS and the look-ahead strategy largely increases the recovery rates for model C, where standard meta-CART algorithms suffer from the local minimum problem and instability. For RE meta-CART, applying SSS and the look-ahead strategy also results in higher full recovery rates for models B, D and E. However, for FE meta-CART, applying SSS and the look-ahead strategy does not improve the recovery rates for models B and E. For model D, it seems that applying SSS and the look-ahead strategy may slightly decrease the recovery rate of FE meta-CART. The recovery performance of RE meta-CART with SSS and the look-ahead strategy is satisfactory (i.e., full recovery rates are larger than 0.80) for models B and C when the number of studies is large enough ($K = 120$). For models D and E, where the underlying structures are difficult to be fully captured by tree models, the partial recovery rate of RE meta-CART with SSS and look-ahead is higher than 0.60 for cases with large sample size ($K = 120$) or no residual heterogeneity ($\sigma_\tau^2 = 0$).

Figure 6.3 presents parallel boxplots of the prediction errors of the four implementations of meta-CART compared with the null model (i.e., predicting the study effect sizes as the overall effect size in the training set) when $\sigma_\tau^2 = 0.025$. Again, RE meta-CART applying SSS with look-ahead has the best performance in most scenarios, and the prediction performance of model C is largely improved by applying SSS with look-ahead. The results are similar when $\sigma_\tau^2 = 0$ or 0.05.

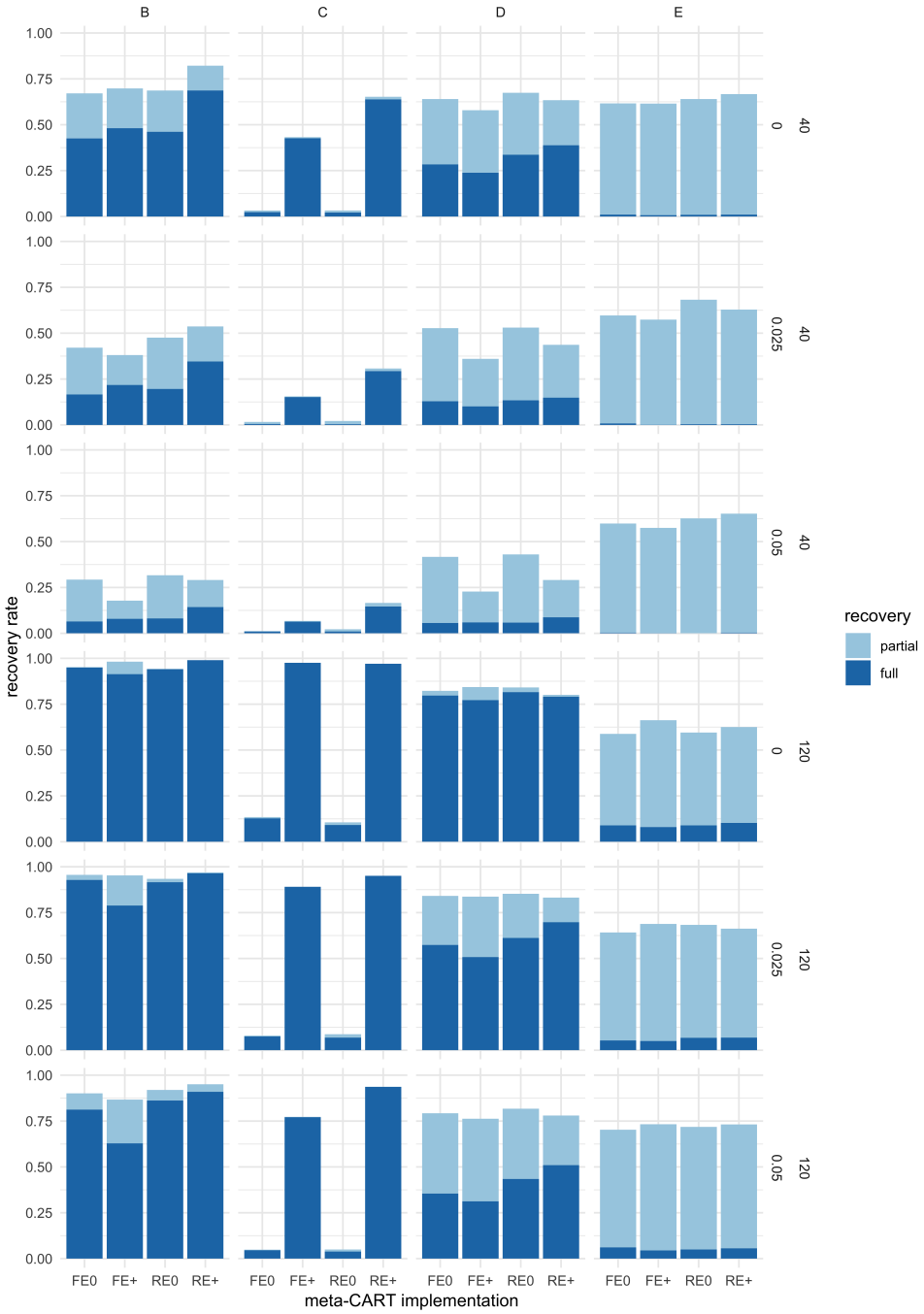


Figure 6.2: The recovery rate of four implementations of meta-CART algorithms: the standard FE meta-CART (FE0), FE meta-CART applying SSS with look-ahead (FE+), the standard RE meta-CART (RE0), and RE meta-CART applying SSS with look-ahead (RE+).

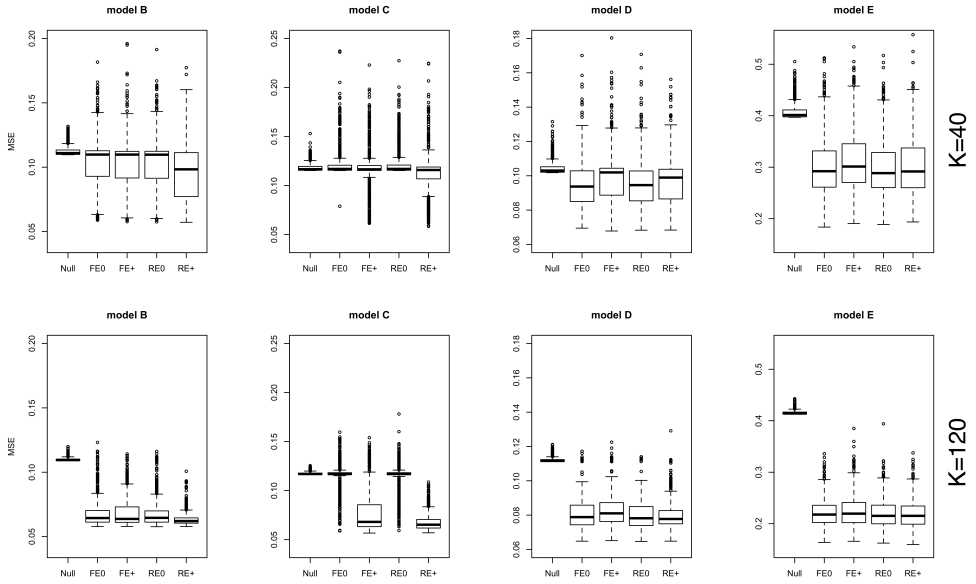


Figure 6.3: The prediction errors of the null model (Null) and the four implementations of meta-CART algorithms: the standard FE meta-CART (FE0), FE meta-CART applying SSS with look-ahead (FE+), the standard RE meta-CART (RE0), and RE meta-CART applying SSS with look-ahead (RE+). The y-axes of the plots for the various models differ due to the difference in the range of the prediction errors.

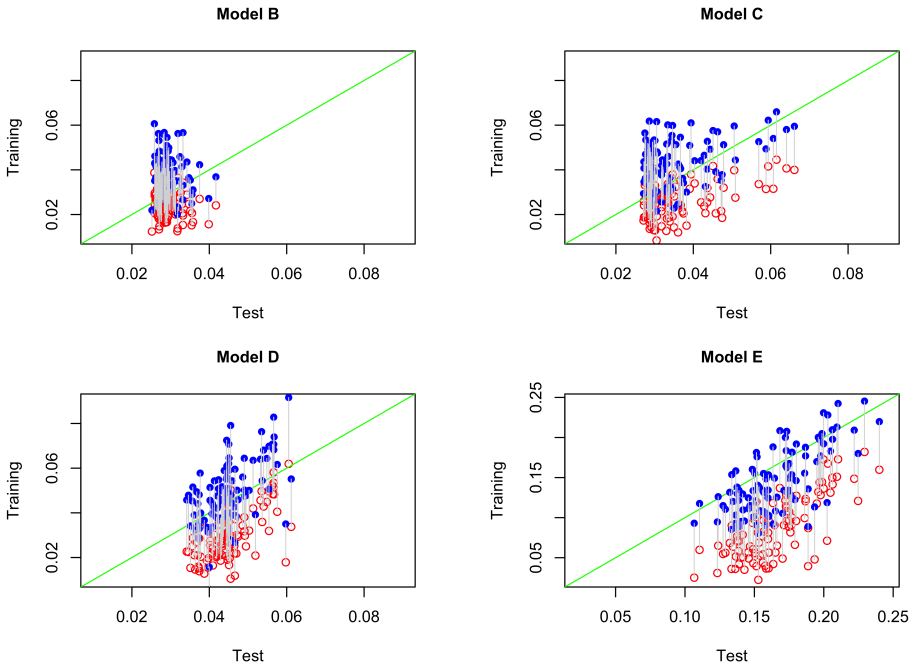


Figure 6.4: Plots of the residual heterogeneity estimates in the training samples (vertical axis) versus the test samples (horizontal axis). The axes of the various models differ due to the difference in the estimated residual heterogeneity. The bias-corrected (blue dots) and uncorrected (red circle) estimates of the residual heterogeneity for the same study are connected by a gray line. The reference line (in green) is $y = x$.

Table 6.2: The false positive rates of the four implementations of meta-CART: the standard FE meta-CART (FE0), FE meta-CART applying SSS with look-ahead (FE+), the standard RE meta-CART (RE0), and RE meta-CART applying SSS with look-ahead (RE+). The rates of identifying non-trivial trees, the rates of non-trivial trees with significant moderator effects tested by the Q-test (p-value < 0.05), and the rates of non-trivial trees with significant moderator effects tested by the permutation test are displayed.

K	σ_τ^2	FE0			FE+		
		non-trivial trees	Q-test	permutation	non-trivial trees	Q-test	permutation
40	0	.046	.046	.026	.016	.016	.007
	0.025	.046	.046	.027	.024	.024	.011
	0.05	.058	.058	.029	.013	.013	.008
120	0	.011	.011	.011	.002	.002	.002
	0.025	.021	.021	.019	.004	.004	.001
	0.05	.023	.023	.017	.002	.002	.002
		RE0			RE+		
40	0	.081	.079	.031	.050	.043	.015
	0.025	.081	.081	.031	.051	.046	.018
	0.05	.077	.077	.031	.054	.047	.016
120	0	.030	.030	.018	.019	.017	.005
	0.025	.053	.053	.028	.030	.025	.011
	0.05	.040	.040	.024	.019	.015	.007

6.5.2 Permutation Test

Table 6.2 presents the false positive rates before and after correction by the permutation test. It shows that the Q-test without permutation test correction has little influence on the control of the false positive rate. Once a non-trivial tree is identified after the cross-validation procedure, the Q-test produces significant results in most cases (p-value < 0.05). In contrast, the permutation test does help with reducing the false positive rates, especially when the number of studies is small ($K = 40$). In addition, when signals do exist, the decrease in the recovery rates after the permutation test correction is small. When $K = 40$, the decrease in the full recovery rate ranges between 0.000 and 0.074 with median = 0.000, and the decrease in the partial recovery rate ranges between 0.000 and 0.082 with median = 0.008 . When $K = 120$, the decrease in the full recovery rate ranges between 0.000 and 0.003 with median = 0.000, and the decrease in the partial recovery rate ranges between 0.000 and 0.011 with median 0.000. Thus, the study results show that the permutation test improves the control of the false positive rate, with little sacrifices in the recovery rate.

6.5.3 Bootstrap Correction

Since the bootstrap correction is only necessary for the RE model (see section 6.3.4), and RE meta-CART applying SSS and the look-ahead strategy has better recovery performance than the standard RE meta-CART, we performed the bootstrap correction only for the subgroup analysis results obtained by RE meta-CART applying SSS and the look-ahead. Figure 6.4 shows the estimated residual heterogeneity in the test sets versus the estimated residual heterogeneity in the training sets when $K = 120, \sigma_\tau^2 = 0.025$. It is shown that the uncorrected residual heterogeneity estimates (i.e., the red circles) are systematically underestimated in the training sets (i.e., the circles are all below the green line). After bias correction, the residual heterogeneity estimates (i.e., the blue dots) become reasonably close in most cases to the estimated values in the test sets (i.e., closer to the green line). The results are similar for the other cells of the simulation design.

6.6 An Illustrative Application

For further illustration of the extended meta-CART, we apply the method to the meta-analytic data collected by Levine et al. (2017). The data consists of 244 samples from 185 studies that report human sperm count. As some studies report more than one sample, the number of samples $K = 244$ is larger than the number of studies. In this application, we are interested in whether the sperm concentration (i.e., the number of sperm cells per micro liter in a semen sample) is influenced by moderators such as time period, fertility status and geographic group, and whether interactions between these moderators play a role. RE meta-CART with the SSS and look-ahead strategy has been applied to the data using the sperm count measure as the effect size, and three moderators: the midpoint of the sample collection period, a binary variable indicating whether the sample has been selected by fertility status (i.e., "Fertile" samples are men that were known to have conceived a pregnancy, and "Unselect" samples are men not selected by fertility status), and a binary variable indicating whether the sample was collected from Western countries such as North America, Europe, Australia, and New Zealand or not. A total of $P = 1000$ samples have been used for the permutation test to assess the significance of the moderator effects. The confidence intervals of the subgroup estimates are computed with bias correction, using $B = 50$ bootstrap samples.

The analysis identifies three subgroups characterized by two moderators: geographic group and time period (see Figure 6.5). The first subgroup consists of 87 studies that are collected from Western countries, and the midpoints of the sample collection period are earlier than 1990. The weighted mean of sperm concentration is highest in this subgroup ($\bar{g} = 90.63$, CI: [86.69, 94.56]). The second subgroup consists of 88 studies that are collected from Western countries, and the midpoints of the sample collection

period are later than 1990. This subgroup has the lowest sperm concentration ($\bar{g} = 70.76$, CI: [67.14, 74.39]). The third subgroup consists of 69 studies that are collected from non-Western countries. The weighted average sperm concentration of this subgroup is in between the other two subgroups ($\bar{g} = 76.53$, CI: [72.33, 80.73]). The permutation test indicates significant heterogeneity between these three subgroups ($Q_b^* = 61.371$, p-value < 0.001). To summarize the analysis results, the sperm concentration is associated with two moderators: the geographic group of the samples, and the time period when the samples were collected. The interaction effect between geographic group and time period reveals that there is a decline of sperm concentration over time in the Western countries, but not in the non-Western countries. These findings are similar to the meta-regression analysis results reported in Levine et al. (2017).

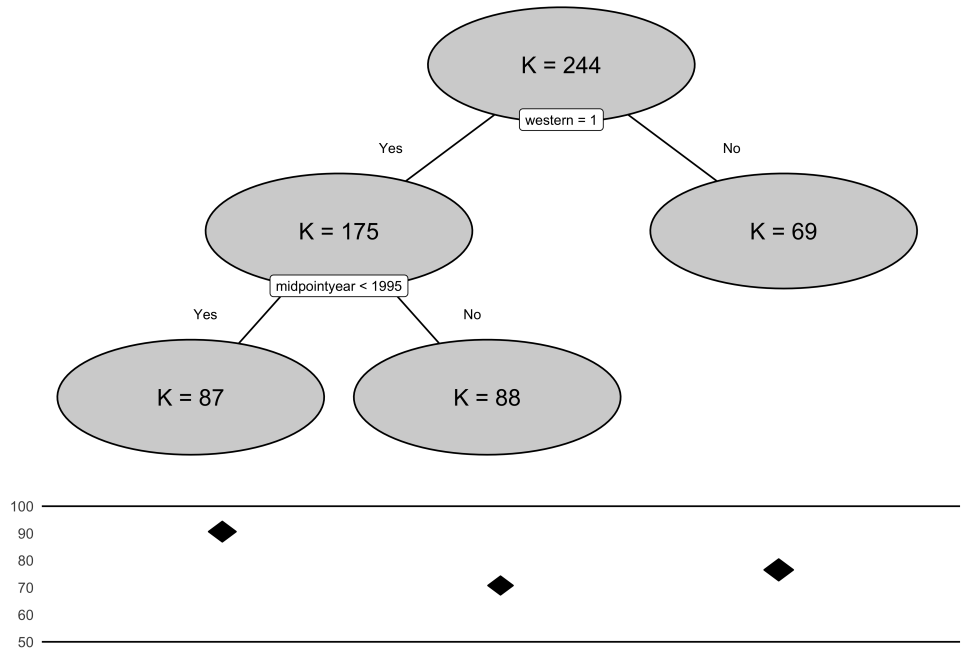


Figure 6.5: The RE meta-CART analysis results on the meta-analytic data collected by Levine et al. (2017). Three subgroups are defined by two splits: 1) whether the sample is collected from Western countries (`werstern` = 1), and 2) whether the midpoint of the data collection period is earlier than 1995 (`midpointyear` < 1995). The height of the diamonds between the solid lines represents the 95% confidence intervals of the weighted subgroup means of sperm concentration.

.

6.7 Discussion

The meta-CART methodology identifies subgroups of homogeneous studies by searching for influential moderators that can explain the heterogeneity in effect sizes. In this paper, we have proposed several extensions for meta-CART to tackle the problems of instability and over-optimism in the subgroup analysis results. Sigmoid smooth surrogate (SSS)

and look-ahead strategies are introduced to alleviate the local optimum problem resulting from the greedy search procedure and to improve the stability of the tree-based model. A permutation test has been developed as an appropriate significance test of the heterogeneity between the identified subgroups. We propose a special bootstrap procedure for meta-CART to correct for the over-optimism in the estimates of the residual heterogeneity unexplained by the moderators, and thereby to correct for the over-optimistic confidence intervals of the subgroup effect sizes. The performance of these extensions are evaluated in extensive simulation studies. These simulation results show that applying SSS and the look-ahead strategy improves the performance of random effects (RE) meta-CART by reducing the false positive rate and improving the recovery rate, especially in the scenario where standard meta-CART suffers from the local optimum problem. Also, the permutation test improves the control of false positive findings with little sacrifice in recovery rates. Finally, the bootstrap procedure works well in correcting for the bias in the estimates of the residual heterogeneity.

Regarding the improvement of the recovery performance, it is no surprise that the combination of SSS and look-ahead works better for RE meta-CART than for FE meta-CART. Because the tree-growing procedure of RE meta-CART is a sequential partitioning algorithm, compared to the recursive partitioning algorithm of FE meta-CART, it is more sensitive to non-optimal splits at the top levels, and thus is more likely to encounter the instability problem. It is also worth to mention that RE meta-CART outperforms FE meta-CART in most scenarios, even in cases of no residual heterogeneity. The reason could be that residual heterogeneity exists unless all the sources of heterogeneity are identified. Thus, the FE assumption is seldom satisfied in the tree construction process.

The strength of the simulation studies is that the models used to generate data sets include both models that can be expressed by trees and models that are difficult for trees to fully recover. One limitation of the simulation studies is that due to the computational cost the look-ahead strategy was applied only to the first two splits. Another limitation is that the simulation studies only focused on continuous moderator variables, because the mechanism of dealing with ordinal or categorical moderators is very similar to dealing with continuous moderators. In addition, Li et al. (2019) showed that standard meta-CART did not perform well for continuous moderators. Therefore, we focused in this study on the improvement of the method with regard to continuous moderators.

An alternative solution for solving the local optimum problem is so-called bumping. Bumping is a bootstrap-based method that can be used to find a better local minimum while preserving the structure of the estimator (Tibshirani & Knight, 1999). The result of bumping is only a single tree, which preserves the interpretatbility of the tree structure. However, the main difficulty of applying bumping is that the tree complexity parameter needs to be pre-defined so that the trees built from different bootstrap samples are of the same complexity and can be compared. The prior knowledge of the tree complexity

parameter is usually unavailable in practice. But if domain knowledge is available to pre-define the number of subgroups, bumping can be another attractive tool to avoid the local optimum problem.

The contribution of the permutation test is not only to reduce the false positive rate, but also to produce interpretable test results. The between-subgroups Q-test used in the standard meta-CART is a pseudo test due to its post-hoc property. Thus, it only gives information if the analysis results are not significant, and the resulted p-value cannot be interpreted as face value. In contrast, the permutation test serves as a more appropriate significance test for the between-subgroup heterogeneity, and the resulting p-value is interpretable.

The idea underlying the bootstrap correction for meta-CART is comparable to the bootstrap technique proposed by Loh et al. (2015) to construct confidence intervals for tree-based methods in a single study setting. Both techniques construct trees using the original data and bootstrap samples of the data, and estimate the within-node variance/heterogeneity by sending down the observations in a node from the original tree into different nodes of the other trees constructed from the bootstrap samples. The main difference is that in the framework of meta-analysis, the observations (i.e., the studies) need to be weighted by the study accuracy. Another difference is that in a single study setting, the correction for confidence intervals is achieved directly by correcting the within-node variance, whereas in the framework of meta-analysis, the correction for confidence intervals is achieved by first correcting for the estimate of the residual heterogeneity.

To summarize, in this paper we proposed four extensions in the framework of meta-CART analysis, and evaluated their performance. We showed that applying both the smooth sigmoid surrogate method and the look-ahead strategy improves the recovery performance of the meta-CART method. In addition, permutation test and bootstrap correction work well in correcting for the over-optimism in the significance test and the confidence intervals of the subgroup effect sizes, respectively. For future work, we suggest that these extensions are also applied to other tree-based methods that aim for subgroup identification (e.g., Dusseldorp & Van Mechelen, 2014; Loh, 2002). In practice, we recommend to use RE meta-CART in favor of FE meta-CART, and to apply both SSS and the look-ahead strategy in the tree construction process. The permutation test and the bootstrap correction are recommended to be applied for statistical inference with respect to the identified subgroups.

Chapter 7

Epilogue

In meta-analysis, heterogeneity often exists between studies, and it is important to identify the sources of heterogeneity when summarizing the study results. Meta-CART is a useful tool to explain seemingly contradictory findings. By identifying the influential study characteristics (i.e., moderators), meta-CART partitions the studies into more homogeneous subgroups with regard to the effect size, and gives a clearer picture for the synthesis of research findings. This dissertation focuses on the developments of the meta-CART method, and aims at identifying homogeneous subgroups of studies in meta-analysis using the moderators.

In Chapter 2, we extended the original meta-CART proposed by (Dusseldorp et al., 2014) and investigated the performance of the extended options via an extensive simulation study. The simulation study showed that meta-regression trees instead of meta-classification trees should be used for a good control of Type I error rate and better performance in retrieving the true structure underlying the data (i.e., recovery performance). Therefore, the subsequent developments of meta-CART are based only on meta-regression trees. In Chapter 3, in order to use the fixed or random effects assumption consistently in both splitting and test procedures, two separate algorithms were proposed for fixed effect (FE) and random effects (RE) meta-CART. The FE meta-CART algorithm is implemented by applying FE weights (i.e, the inverse of sampling variance) in the CART algorithm. The RE meta-CART uses a sequential partitioning algorithm that updates the estimate of the residual heterogeneity each time when a new split is introduced. In Chapter 4, an R-package was introduced to implement the meta-CART algorithms in the R language. Also, a look-ahead strategy for the greedy search was proposed to improve the performance of meta-CART. In Chapter 6, a combination of the smooth sigmoid surrogate (SSS) and look-ahead strategies was proposed to alleviate the instability and local optimum issues in the splitting procedure. In addition, permutation test and bootstrapping were applied to correct the over-optimism in the test procedure.

The simulation studies in Chapters 2 and 3 examined the performance of meta-CART on data generated from models that can be expressed by trees. The results showed that

the performance of meta-CART depends on various design factors. In summary, the performance of meta-CART is positively related to the number of studies, the within-study sample size, and the magnitude of interaction effect size. On the other hand, the performance of meta-CART is negatively influenced by the number of moderators, the residual heterogeneity, and the complexity underlying the data. The simulation study in Chapter 6 investigated the performance of the extended meta-CART implementations on data generated from more complex models, including both models that can be expressed by trees and models that cannot be fully captured by trees. The results showed that in situations that are difficult for meta-CART to fully retrieve the underlying structure, the method is still able to identify (part of) the true moderators with no spurious findings. In addition, when there is a severe local optimum problem, applying the SSS and look-ahead strategies can largely improve the performance of meta-CART.

Based on the simulation results, guidelines for the practical use of meta-CART have been formulated. RE meta-CART should be used in favor of FE meta-CART, because the RE assumption takes into account the residual heterogeneity unexplained by the identified moderators, and this is more realistic in practice. As mentioned above, we recommend using SSS and look-ahead strategies in the tree construction process to alleviate the local optimum problem. Also, to correct for possible over-optimism in the subgroup analysis results, a permutation test and bootstrap correction are recommended to be applied. Regarding the pruning rule, the half-standard-error rule results in a better balance between Type I error and power compared to the minimum-standard-error rule and the one-standard error rule. To perform a meta-CART analysis with satisfactory performance (i.e., with power and recovery rates both higher than 0.80), a minimum number of 40 studies is required to detect main effects or simple interaction effects such as a two-way interaction, and a minimum number of 80 studies is required for the identification of more complex interaction effects (e.g., multiple two-way interactions or higher-order interaction).

The applications of meta-CART have been illustrated by several real-world data sets in Chapters 4, 5 and 6. Different measures of effect size were used in these meta-analytic studies, such as standardized mean difference, log-odds ratio, and counts. In general, meta-CART can be applied to any measure of effect size that satisfies the Central Limit Theorem. That is, the distribution of the effect size measure should approach normality eventually for large enough within-study sample size, as the effect size is commonly measured by sample averages. If the distribution of the effect size measure is highly-skewed, transformation can be used. For example, Fisher's Z transformation is used for correlation coefficient (see Fisher, 1915).

In single study settings, tree-based methods are popular tools for prediction, subgroup identification, and explanatory purposes. However, in the framework of meta-analysis, there are several difficulties in applying meta-CART in practice. First, studies in a meta-

analysis are often restricted to published studies that satisfy certain inclusion criteria (e.g., English publications, randomized controlled trials). Thus, the sample sizes of meta-analyses (i.e., the number of studies) are often small. In such cases, meta-CART analysis can be under-powered. As mentioned above, we learned that meta-CART requires at least 40 studies for identification of simple moderator effects and 80 studies for identification of complex moderator effects to achieve good performance. However, in practice the prior information about whether the moderator effects are complex or not is usually unavailable. This makes it difficult for researchers to decide whether the sample size is large enough to conduct a meta-CART analysis. In addition, unlike the data collection process in single study settings, the number of studies in a meta-analysis is not known until the collection process is completed. This also makes it difficult for researchers to decide whether to apply a meta-CART analysis when they pre-register for a meta-analysis. Second, there are often missing values in the potential moderator variables. Since meta-analytic data are usually collected from independent studies, the interested potential moderators may not have been reported in all studies. Although tree-based methods can deal with missing values easily by treating the missing values as a new class "missing" for categorical variable, and randomly assigning the studies with missing values into the children nodes for continuous variable, a large number of missing values can deteriorate the performance of meta-CART by decreasing the stability and the power of the analysis. A possible solution for this problem could be using common guidelines in certain fields for researchers to report their scientific findings with a list of well-defined interested moderators variables. For example, in the field of health psychology, Abraham and Michie (2008) proposed a taxonomy of behavior change techniques (BCT) used in health psychological interventions. This taxonomy is used to improve the specification of behavior change interventions, and was further refined by Michie et al. (2011) for interventions that aim to support people who change their physical activity and healthy eating behaviors. The taxonomy has been used by many health psychologist when reporting their research findings. This makes it easier for researchers who conduct meta-analyses to synthesize these research findings by coding the BCTs that were used in different interventions as potential moderators (for an example of BCT meta-analyses, see van Genugten et al., 2016). Third, it is often argued that publication bias exists in many fields (Rothstein, Sutton, & Borenstein, 2005), and such bias in the primary studies will lead to biased results in meta-analysis (Borenstein et al., 2009). Publication bias refers to the phenomenon that studies with non-significant effect sizes have less probability to get published than those with significant effect sizes. For a standard meta-analysis, van Aert, Wicherts, and van Assen (2016); Van Assen, van Aert, and Wicherts (2015) proposed to use the p-uniform and p-curve methods to correct for publication bias. How to correct for the publication bias in meta-CART analysis remains a challenging question, and this issue has not been addressed in this thesis. One possible direction is to fit a tree including both potential moderators and the variables

that could possibly be related to the publication bias, such as within-study sample size, or dummy variables that indicate whether the study effect size is significant or not. It would be interesting in future work to investigate the influence of publication bias on moderator analysis, especially moderator analysis by meta-CART, and to evaluate whether including additional variables in tree modeling can improve the detection or correction for publication bias.

Another interesting avenue for future research would be the application of the proposed extensions in other tree based methods. Although this thesis mainly focuses on the meta-CART method, some of the proposed extensions can be also applied to other single-tree-based methods that aim at interpretable results. The SSS strategy and look ahead strategies can be applied to other greedy search methods to alleviate the local optimum problem, and the permutation test and the bootstrapping correction can be applied to draw statistical inferences in other subgroup identification methods. For example, it might be interesting to apply these extensions to the IPD-SIDES proposed by Mistry et al. (2018), which integrates the subgroup identification based on different effect search (SIDES; Lipkovich et al., 2011) into the framework of individual patient data (IPD) meta-analysis.

In conclusion, meta-CART is a promising alternative of meta-regression to explain the heterogeneity between studies by exploring the relationship between moderator variables and study effect sizes. The potential of applying tree-based methods in meta-analysis has not been fully explored. This is partly due to the facts that the researchers are seldom in a data-rich situation when conducting a meta-analysis, and the quality of the data highly depends on the included primary studies. In this thesis, several extensions have been proposed to overcome the limitations from the methodology aspect, and the performance of the extended meta-CART implementations has been evaluated. The meta-CART method is able to identify homogeneous subgroups of studies, by searching for the moderators that influence study effect sizes. It can deal with various types of moderators, including dichotomous, nominal, ordinal and continuous variables. The heterogeneity between subgroups is tested by the between-subgroups Q-statistic, and the uncertainty in the estimates for the subgroup effect sizes are quantified by confidence intervals. I hope that my work will increase other researchers' interest in using tree-based methods in the field of meta-analysis, and that the proposed extensions can also be applied to other tree-based methods in future. Furthermore, although the data collection process in meta-analysis is not a focus in this thesis, I hope that there would be a larger awareness of how to improve the quality of meta-analytic data by improving the quality of reporting primary publications.

Chapter 8

Appendix

8.1 The SMD effect size in meta-analysis

In studies that compare treatment and control groups with respect to some continuous response variable, standardized mean difference (SMD) is commonly chosen as the measure of effect size. Several measures are available to compute the SMD, of which the most popular one is Cohen's d (Cohen, 1988). It is computed as the mean difference between the treatment and control group divided by the pooled standard deviation leading to the estimator:

$$d_k = \frac{\overline{Y}_k^T - \overline{Y}_k^C}{S_k}, \tag{8.1}$$

where $\overline{Y}_k^T$ and $\overline{Y}_k^C$ denote, respectively, the mean of the treatment group and control group of the k^{th} study, for $k = 1, \ldots, K$, with K being the total number of studies in the meta-analysis. S_k is the pooled standard deviation of the k^{th} study. This standard deviation is computed as

$$S_k = \sqrt{\frac{(n_k^T - 1)(S_k^T)^2 + (n_k^C - 1)(S_k^C)^2}{n_k^T + n_k^C - 2}}, \tag{8.2}$$

where n_k^T, n_k^T, S_k^T, and S_k^T are, respectively, the treatment and control group sample sizes and standard deviations of the kth study (also see Cohen (1988)).

When sample sizes of the individual studies are not sufficiently large, the SMDs are positively biased. Therefore, Hedges (1981) introduced a modified estimator for the SMD with a correction for small sample size. This modified estimator is sometimes referred as Hedges' g (Hedges, 1981), and it is given by

$$g_k = d_k \cdot c(m_k), \tag{8.3}$$

where $m_k = n_k^T + n_k^C - 2$. $c(m_k)$ only depends on m_k and can be computed by

$$c(m_k) = \frac{\Gamma(m_k/2)}{\sqrt{m_k/2}\Gamma((m_k-1)/2)}. \tag{8.4}$$

The constant $c(m_k)$ is less than unity and approaches unity when m_k is large. It can be closely approximated by

$$c(m_k) \approx 1 - \frac{3}{4m_k - 1}. \tag{8.5}$$

The SMD is not normally distributed. In fact, it is closely related to a non-central t-distribution

$$c(m_k)^{-1}(\tilde{n_k})^{1/2} g_k \sim t_{m_k}(\delta_k \sqrt{\tilde{n_k}}), \tag{8.6}$$

where δ_k is the true effect size of the k^{th} study, and $\tilde{n_k} = (n_k^T n_k^C)/(n_k^T + n_k^C)$.

8.2 Test of Heterogeneity

8.2.1 Testing heterogeneity of effect sizes across studies

Denote the true effect size in the k^{th} study by δ_k. And denote the sampling variance in the k^{th} study by $\sigma^2_{\epsilon_k}$. The observed effect size in the kth study g_k is distributed as $\mathcal{N}(\delta_k, \sigma^2_{\epsilon_k})$. The sampling variance $\sigma^2_{\epsilon_k}$ depends on the within-study sample sizes and the effect size of the k^{th} study. It can be estimated by

$$\hat{\sigma}^2_{\epsilon_k} = \frac{n_k^T + n^C}{n_k^T n_k^C} + \frac{g_k^2}{2(n_k^T + n_k^C)}. \tag{8.7}$$

Under the null hypothesis $\delta_1 = \delta_2 = ... = \delta_K$, the weighted estimator of effect size is given by $g_+ = \sum_{k=1}^K w_k g_k$, where w_k is computed by

$$w_k = \frac{1}{\hat{\sigma}^2_{\epsilon_k}} / \sum_{k=1}^K \frac{1}{\hat{\sigma}^2_{\epsilon_k}}. \tag{8.8}$$

Then the test statistic

$$Q = \sum_{k=1}^K \frac{(g_k - g_+)^2}{\hat{\sigma}^2_{\epsilon_k}} \tag{8.9}$$

has an asymptotic chi-square distribution with $K-1$ degrees of freedom. The null hypothesis will be rejected if the value of Q exceeds the $100(1-\alpha)$th percentile of the chi-square distribution with $K-1$ degrees of freedom (see in Hedges, 1981).

8.2.2 Testing heterogeneity of effect sizes between subgroups

For subgroup meta-analysis, the generally advocated model is a mixed effects model (Borenstein et al., 2009). In a mixed effects meta-analysis, the within-subgroup effect

sizes are computed by using the random effects assumption. Then a fixed effects model is employed to test the heterogeneity between subgroups. Denote the effect size of the k^{th} study in the b^{th} subgroup by g_{bk}, the summary effect size of the b^{th} subgroup can be computed as

$$g_{b+} = \sum_{k=1}^{K_b} \frac{g_{bk}}{\hat{\sigma}^2_{\epsilon_{bk}} + \sigma^2_{\tau_b}} / \sum_{k=1}^{K_b} \frac{1}{\hat{\sigma_b}^2_{\epsilon_{bk}} + \sigma^2_{\tau_b}}, \tag{8.10}$$

where K_b is the number of studies in the b^{th} subgroup, and $\hat{\sigma}^2_{\epsilon_{bk}}$ and $\sigma^2_{\tau_b}$ are the sampling variance and the residual heterogeneity (see Section 2). The variance of the summary effect size is estimated as

$$\hat{\sigma}^2_{g_{b+}} = 1 / \sum_{k=1}^{K_b} \frac{1}{\hat{\sigma}^2_{\epsilon_{bk}} + \sigma^2_{\tau_b}}. \tag{8.11}$$

The weighted mean of the within-subgroup summary effect sizes is computed as

$$g_{+} = \sum_{b=1}^{B} \frac{g_{b+}}{\hat{\sigma}^2_{g_{b+}}} / \sum_{b=1}^{B} \frac{1}{\hat{\sigma}^2_{g_{b+}}}, \tag{8.12}$$

where B is the number of subgroups. The between-subgroup Q statistic can be computed as

$$Q_{between} = \sum_{1}^{B} \frac{(g_{b+} - g_{+})^2}{\hat{\sigma}^2_{g_{b+}}}. \tag{8.13}$$

8.3 Partitioning criterion in CART

Starting from one group including all subjects (i.e., root node), CART algorithm partitions the root node into two subgroups (i.e., offspring nodes), by searching all possible splits across all predictor variables to find the split that best satisfies the partitioning criterion. Then the splitting process is repeated on the offspring nodes to grow the tree. There are various impurity function that can be used as partitioning criterion.To generalize, consider a proposed split s of a node t, into two offspring nodes l and r with the proportion p_l of the data cases in t to l and the proportion p_r to r. Let $i(t)$ be some impurity function of node t. Then the decrease in impurity by split s of node t will be

$$\Delta i(s,t) = i(t) - p_l i(l) - p_r i(r). \tag{8.14}$$

Δi can be used as a goodness of split: a high value of $\Delta i(s,t)$ means the proposed split s is good.

For regression tree, the within-node sum of squares is often used as the impurity function. It can be written as

$$i(t) = \frac{1}{N} \sum_{(v_n, z_n) \in t} (z_n - \bar{z}(t)), \tag{8.15}$$

where (v_n, y_n) are the data cases with v_n being the predictor vector and z_n being the response variable; $\bar{z}(t)$ is the mean of z_n for all cases (v_n, z_n) that fall into the node t.
For classification tree, one of the most frequently used impurity function is the Gini index, which can be written as

$$i(t) = \sum_{f \neq h} p(f|t)p(h|t), \tag{8.16}$$

where $p(f|t)$ is the probability that the subject belongs to the f^{th} class given it falls in node t. The Gini index is 0 when subjects from only one same class fall in a node, and it achieves its maximum (0.5 in case of two classes) when each class has the same probability to fall in a node.

Bibliography

Abraham, C., & Michie, S. (2008). A taxonomy of behavior change techniques used in interventions. *Health psychology*, *27*(3), 379.

Arthur Jr, W., Bennett, W., & Huffcutt, A. I. (2001). *Conducting meta-analysis using sas*. Psychology Press.

Avenell, A., Broom, J., Brown, T. J., Poobalan, A., Aucott, L., Stearns, S., & et. al. (2004). Systematic review of the long-term effects and economic consequences of treatments for obesity and implications for health improvement. *Health technology assessment*, *8*(21).

Bartlett, Y. K., Sheeran, P., & Hawley, M. S. (2014). Effective behaviour change techniques in smoking cessation interventions for people with chronic obstructive pulmonary disease: A meta-analysis. *British Journal of Health Psychology*, *19*(1), 181–203.

Bastian, H., Glasziou, P., & Chalmers, I. (2010). Seventy-five trials and eleven systematic reviews a day: how will we ever keep up? *PLoS Medicine*, *7*(9), e1000326.

Borenstein, M., Hedges, L. V., Higgins, J., & Rothstein, H. R. (2010). A basic introduction to fixed-effect and random-effects models for meta-analysis. *Research Synthesis Methods*, *1*(2), 97–111. doi: 10.1002/jrsm.12

Borenstein, M., Hedges, L. V., Higgins, J. P., & Rothstein, H. R. (2009). *Introduction to meta-analysis*. Hoboken, NJ: Wiley. doi: 10.1002/9780470743386

Boulesteix, A.-L. (2006). Maximally selected chi-square statistics and binary splits of nominal variables. *Biometrical Journal*, *48*(5), 838–848.

Bourassa, D., McManus, I. C., & Bryden, M. P. (1996). Handedness and eye-dominance: a meta-analysis of their relationship. *Laterality: Asymmetries of Body, Brain and Cognition*, *1*(1), 5–34. doi: 10.1080/713754206

Boutron, I., Moher, D., Altman, D. G., Schulz, K. F., & Ravaud, P. (2008). Extending the consort statement to randomized trials of nonpharmacologic treatment: explanation and elaboration. *Annals of internal medicine*, *148*(4), 295–309.

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and regression trees*. Wadsworth, Belmont, CA: CRC press.

Brent, R. P. (2013). *Algorithms for minimization without derivatives*. Courier Corporation.

Bull, E. R., Dombrowski, S. U., McCleary, N., & Johnston, M. (2014). Are interventions for low-income groups effective in changing healthy eating, physical activity and smoking behaviours? a systematic review and meta-analysis. *BMJ open*, *4*(11), 006–046.

Bull, E. R., McCleary, N., Li, X., Dombrowski, S. U., Dusseldorp, E., & Johnston, M. (2018). Interventions to promote healthy eating, physical activity and smoking in low-income groups: a systematic review with meta-analysis of behavior change techniques and delivery/context. *International Journal of Behavioral Medicine*, *25*(6), 605–616.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu, C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of clinical epidemiology*, *46*(5), 423–429.

Carver, C. S., & Scheier, M. F. (1982). Control theory: A useful conceptual framework for personality–social, clinical, and health psychology. *Psychological bulletin*, *92*(1), 111.

Chang, M.-W., Nitzke, S., & Brown, R. (2010). Design and outcomes of a mothers in motion behavioral intervention pilot study. *Journal of Nutrition Education and Behavior*, *42*(3), S11–S21.

Chesterman, J., Judge, K., Bauld, L., & Ferguson, J. (2005). How effective are the english smoking treatment services in reaching disadvantaged smokers? *Addiction*, *100*, 36–45.

Cheung, M. W.-L., & Chan, W. (2005). Meta-analytic structural equation modeling: a two-stage approach. *Psychological Methods*, *10*(1), 40. doi: 10.1037/1082-989X.10 .1.40

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Collaboration, T. (2008). Review manager (revman). *Copenhagen: The Nordic Cochrane Centre*.

Corp, I. (2013). Ibm spss statistics for windows, version 22.0 [Computer software manual]. Armonk, NY. Retrieved from `http://www.spss.com/`

Dangour, A. D., Albala, C., Allen, E., Grundy, E., Walker, D. G., Aedo, C., ... Uauy, R. (2011). Effect of a nutrition supplement and physical activity program on pneumonia and walking capacity in chilean older people: a factorial cluster randomized trial. *PLoS medicine*, *8*(4), e1001023.

Davidson, K. W., Goldstein, M., Kaplan, R. M., Kaufmann, P. G., Knatterud, G. L.,

Orleans, C. T., ... Whitlock, E. P. (2003). Evidence-based behavioral medicine: what is it and how do we achieve it? *Annals of behavioral medicine*, *26*(3), 161–171.

DerSimonian, R., & Laird, N. (1986). Meta-analysis in clinical trials. *Controlled Clinical Trials*, *7*(3), 177–188.

Dombrowski, S. U., Sniehotta, F. F., Avenell, A., Johnston, M., MacLennan, G., & Araujo-Soares, V. (2012). Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. *Health Psychology Review*, *6*(1), 7–32.

Dombrowski, S. U., Sniehotta, F. F., Avenell, A., Johnston, M., MacLennan, G., & Araújo-Soares, V. (2012). Identifying active ingredients in complex behavioural interventions for obese adults with obesity-related co-morbidities or additional risk factors for co-morbidities: a systematic review. *Health Psychology Review*, *6*(1), 7–32.

Doyle, P. (1973). The use of automatic interaction detector and similar search procedures. *Operational Research Quarterly*, 465–467. doi: 10.2307/3008131

Drewnowski, A., & Specter, S. E. (2004). Poverty and obesity: the role of energy density and energy costs. *The American journal of clinical nutrition*, *79*(1), 6–16.

Dusseldorp, E., Conversano, C., & Van Os, B. J. (2010). Combining an additive and tree-based regression model simultaneously: Stima. *Journal of Computational and Graphical Statistics*, *19*(3), 514–530. doi: 10.1198/jcgs.2010.06089

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

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

Dutton, G. R., Martin, P. D., Welsch, M. A., & Brantley, P. J. (2007). Promoting physical activity for low-income minority women in primary care. *American Journal of Health Behavior*, *31*(6), 622–631.

Efron, B., & Tibshirani, R. J. (1994). *An introduction to the bootstrap*. CRC press.

Elder, J. P., Ayala, G. X., Campbell, N. R., Arredondo, E. M., Slymen, D. J., Baquero, B., & et. al. (2006). Long-term effects of a communication intervention for spanish-dominant latinas. *American Journal of Preventive Medicine*, *31*(2), 159–166.

Emmons, K. M., Stoddard, A. M., Fletcher, R., Gutheil, C., Suarez, E. G., Lobb, R., ... Bigby, J. A. (2005). Cancer prevention among working class, multiethnic adults: results of the healthy directions–health centers study. *American Journal of Public Health*, *95*(7), 1200–1205.

Esmeir, S., & Markovitch, S. (2007). Anytime learning of decision trees. *Journal of Machine Learning Research*, *8*(May), 891–933.

Finch, W. H., Chang, M., Davis, A. S., Holden, J. E., Rothlisberg, B. A., & McIntosh, D. E. (2011). The prediction of intelligence in preschool children using alternative models to regression. *Behavior Research Methods*, *43*(4), 942–952. doi: 10.3758/s13428-011-0102-z

Fisher, R. A. (1915). Frequency distribution of the values of the correlation coefficient in samples from an indefinitely large population. *Biometrika*, *10*(4), 507–521.

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

Friedman, J. (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, *19*(1), 1–67.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 1189–1232.

Friedman, J., Hastie, T., & Tibshirani, R. (2001). *The elements of statistical learning* (Vol. 1) (No. 10). Springer series in statistics New York, NY, USA:.

Gans, K. M., Risica, P. M., Strolla, L. O., Fournier, L., Kirtania, U., Upegui, D., & et. al. (2009). Effectiveness of different methods for delivering tailored nutrition education to low income, ethnically diverse adults. *International Journal of Behavioral Nutrition and Physical Activity*, *6*(1), 24.

Gardner, B. (2015). A review and analysis of the use of 'habit'in understanding, predicting and influencing health-related behaviour. *Health psychology review*, *9*(3), 277–295.

Gasparrini, A., Armstrong, B., & Kenward, M. G. (2012). Multivariate meta-analysis for non-linear and other multi-parameter associations. *Statistics in Medicine*, *31*(29), 3821–3839.

Gilli, M., Maringer, D., & Schumann, E. (2011). *Numerical methods and optimization in finance*. Academic Press.

Government, S. (2008). *Equally well: report of the ministerial task force on health inequalities.* Scottish Government Edinburgh.

Gruer, L., Hart, C. L., Gordon, D. S., & Watt, G. C. (2009). Effect of tobacco smoking on survival of men and women by social position: a 28 year cohort study. *Bmj*, *338*, b480.

Hart, C. L., Gruer, L., & Watt, G. C. (2011). Cause specific mortality, social position, and obesity among women who had never smoked: 28 year cohort study. *Bmj*, *342*, d3785.

Hedges, L. V. (1981). Distribution theory for glass's estimator of effect size and related estimators. *Journal of Educational and Behavioral Statistics*, *6*(2), 107–128. doi: 10.3102/10769986006002107

Hedges, L. V., & Olkin, I. (1985). *Statistical methods for meta-analysis*. Orlando [etc.]:

Academic Press.

Higgins, J., & 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

Hiscock, R., Judge, K., & Bauld, L. (2010). Social inequalities in quitting smoking: what factors mediate the relationship between socioeconomic position and smoking cessation? *Journal of Public Health*, *33*(1), 39–47.

Hoffmann, T. C., Erueti, C., & Glasziou, P. P. (2013). Poor description of non-pharmacological interventions: analysis of consecutive sample of randomised trials. *Bmj*, *347*, f3755.

Hoffmann, T. C., Glasziou, P. P., Boutron, I., Milne, R., Perera, R., Moher, D., ... et. al. (2014). Better reporting of interventions: template for intervention description and replication (tidier) checklist and guide. *British Medical Journal*, *348*, g1687.

Hothorn, T., Hornik, K., & Zeileis, A. (2006). Unbiased recursive partitioning: A conditional inference framework. *Journal of Computational and Graphical statistics*, *15*(3), 651–674.

Huisman, S. D., De Gucht, V., Dusseldorp, E., & Maes, S. (2009). The effect of weight reduction interventions for persons with type 2 diabetes: a meta-analysis from a self-regulation perspective. *The Diabetes Educator*, *35*(5), 818 – 835. doi: 10.1177/0145721709340929

Hunter, J. E., & Schmidt, F. L. (1990). Dichotomization of continuous variables: The implications for meta-analysis. *Journal of Applied Psychology*, *75*(3), 334. doi: 10.1037/0021-9010.75.3.334

Inc, S. I. (2002). Sas/stat software, version 9.1 [Computer software manual]. Carry, NC. Retrieved from `http://www.sas.com/`

Jackson, R. A., Stotland, N. E., Caughey, A. B., & Gerbert, B. (2011). Improving diet and exercise in pregnancy with video doctor counseling: a randomized trial. *Patient education and counseling*, *83*(2), 203–209.

Keppel, G. (1991). Design and analysis. engelwood cliffs. *NJ: Prentic Hall*.

Keyserling, T. C., Hodge, C. D. S., Jilcott, S. B., Johnston, L. F., Garcia, B. A., Gizlice, Z., ... others (2008). Randomized trial of a clinic-based, community-supported, lifestyle intervention to improve physical activity and diet: the north carolina enhanced wisewoman project. *Preventive medicine*, *46*(6), 499–510.

Kushner, R. F. (2007). Obesity management. *Gastroenterology Clinics of North America*, *36*(1), 191–210.

Leach, H. J., O'Connor, D. P., Simpson, R. J., Rifai, H. S., Mama, S. K., & Lee, R. E. (2016). An exploratory decision tree analysis to predict cardiovascular disease risk in african american women. *Health Psychology*, *35*(4), 397. doi: 10.1037/hea0000267

LeBlanc, M., & Crowley, J. (1993). Survival trees by goodness of split. *Journal of the American Statistical Association*, *88*(422), 457–467.

Levine, H., Jørgensen, N., Martino-Andrade, A., Mendiola, J., Weksler-Derri, D., Mindlis, I., ... Swan, S. H. (2017). Temporal trends in sperm count: a systematic review and meta-regression analysis. *Human reproduction update*, *23*(6), 646–659.

Li, X., Dusseldorp, E., Liu, K., & Meulman, J. (2017). Meta-cart: A flexible approach to identify moderators in meta-analysis. [Computer software manual]. (R package version 1.1-2)

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*, *70*(1), 118–136. doi: 10.1111/bmsp.12088

Li, X., Dusseldorp, E., & Meulman, J. J. (2019). A flexible approach to identify interaction effects between moderators in meta-analysis. *Research synthesis methods*, *10*(1), 134–152. doi: 10.1002/jrsm.1334

Lipkovich, I., Dmitrienko, A., & B D'Agostino Sr, R. (2017). Tutorial in biostatistics: data-driven subgroup identification and analysis in clinical trials. *Statistics in Medicine*, *36*(1), 136–196.

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

Lipsey, M. W., & Wilson, D. B. (2001). *Practical meta-analysis* (Vol. 49). Sage publications Thousand Oaks, CA.

Little, T. D. (2013). *The oxford handbook of quantitative methods, volume 2: Statistical analysis*. Oxford University Press.

Loh, W.-Y. (2002). Regression tress with unbiased variable selection and interaction detection. *Statistica Sinica*, 361–386. doi: 10.1080/03610918.2012.674600

Loh, W.-Y., He, X., & Man, M. (2015). A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine*, *34*(11), 1818–1833.

Lorencatto, F., West, R., & Michie, S. (2012). Specifying evidence-based behavior change techniques to aid smoking cessation in pregnancy. *Nicotine & Tobacco Research*, *14*(9), 1019–1026.

Lorencatto, F., West, R., Stavri, Z., & Michie, S. (2012). How well is intervention content described in published reports of smoking cessation interventions? *nicotine & tobacco research*, *15*(7), 1273–1282.

Lumley, T. (2012). rmeta: Meta-analysis [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rmeta` (R package version 2.16)

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the national cancer institute*, *22*(4), 719–748.

Marcus, B. H., Dunsiger, S. I., Pekmezi, D. W., Larsen, B. A., Bock, B. C., Gans, K. M.,

... Tilkemeier, P. (2013). The seamos saludables study: A randomized controlled physical activity trial of latinas. *American journal of preventive medicine*, *45*(5), 598–605.

Maxwell, S. E., & Delaney, H. D. (1993). Bivariate median splits and spurious statistical significance. *Psychological Bulletin*, *113*(1), 181.

McCleary, N., Duncan, E. M., Stewart, F., & Francis, J. J. (2013). Active ingredients are reported more often for pharmacologic than non-pharmacologic interventions: an illustrative review of reporting practices in titles and abstracts. *Trials*, *14*(1), 146.

Merkle, E. C., & Shaffer, V. A. (2011). Binary recursive partitioning: Background, methods, and application to psychology. *British Journal of Mathematical and Statistical Psychology*, *64*(1), 161–181. doi: 10.1348/000711010X503129

Michie, S., Abraham, C., Whittington, C., McAteer, J., & Gupta, S. (2009). Effective techniques in healthy eating and physical activity interventions: a meta-regression. *Health Psychology*, *28*(6), 690 – 701. doi: 10.1037/a0016136

Michie, S., Ashford, S., Sniehotta, F. F., Dombrowski, S. U., Bishop, A., & French, D. P. (2011). A refined taxonomy of behaviour change techniques to help people change their physical activity and healthy eating behaviours: the calo-re taxonomy. *Psychology & health*, *26*(11), 1479–1498.

Michie, S., Jochelson, K., Markham, W. A., & Bridle, C. (2009). Low-income groups and behaviour change interventions: a review of intervention content, effectiveness and theoretical frameworks. *Journal of Epidemiology & Community Health*, *63*(8), 610–622.

Michie, S., Johnson, B. T., & Johnston, M. (2015). Advancing cumulative evidence on behaviour change techniques and interventions: a comment on peters, de bruin, and crutzen. *Health Psychology Review*, *9*(1), 25–29. doi: 10.1080/17437199.2014.912538

Michie, S., Richardson, M., Johnston, M., Abraham, C., Francis, J., Hardeman, W., ... Wood, C. E. (2013). The behavior change technique taxonomy (v1) of 93 hierarchically clustered techniques: building an international consensus for the reporting of behavior change interventions. *Annals of behavioral medicine*, *46*(1), 81–95.

Mistry, D., Stallard, N., & Underwood, M. (2018). A recursive partitioning approach for subgroup identification in individual patient data meta-analysis. *Statistics in Medicine*, *37*(9), 1550–1561.

Morgan, J. N., & Sonquist, J. A. (1963). Problems in the analysis of survey data, and a proposal. *Journal of the American statistical association*, *58*(302), 415–434.

Moyer, C. A., Rounds, J., & Hannum, J. W. (2004). A meta-analysis of massage therapy research. *Psychological bulletin*, *130*(1), 3.

Niederdeppe, J., Fiore, M. C., Baker, T. B., & Smith, S. S. (2008). Smoking-cessation media campaigns and their effectiveness among socioeconomically advantaged and

disadvantaged populations. *American Journal of Public Health*, *98*(5), 916–924.

Norman, G. J., Zabinski, M. F., Adams, M. A., Rosenberg, D. E., Yaroch, A. L., & Atienza, A. A. (2007). A review of ehealth interventions for physical activity and dietary behavior change. *American journal of preventive medicine*, *33*(4), 336–345.

Normand, S.-L. T. (1999). Tutorial in biostatistics meta-analysis: formulating, evaluating, combining, and reporting. *Statistics in Medicine*, *18*(3), 321–359.

of Clinical Excellence., N. I. (2014). *Behavior change: individual approaches (ph 49).* Retrieved from `https://www.nice.org.uk/guidance/ph49`

Olejnik, S., & Algina, J. (2003). Generalized eta and omega squared statistics: measures of effect size for some common research designs. *Psychological methods*, *8*(4), 434.

Olvera, N., Bush, J. A., Sharma, S. V., Knox, B. B., Scherer, R. L., & Butte, N. F. (2010). Bounce: a community-based mother–daughter healthy lifestyle intervention for low-income latino families. *Obesity*, *18*(S1), S102–S104.

Orwin, R. G., & Vevea, J. L. (2009). Evaluating coding decisions. *The handbook of research synthesis and meta-analysis*, *2*, 177–203.

O'Brien, N., McDonald, S., Araújo-Soares, V., Lara, J., Errington, L., Godfrey, A., & et. al. (2015). The features of interventions associated with long-term effectiveness of physical activity interventions in adults aged 55–70 years: a systematic review and meta-analysis. *Health psychology review*, *9*(4), 417–433. doi: 10.1080/17437199.2015.1012177

Peters, G.-J. Y., Ruiter, R. A., & Kok, G. (2013). Threatening communication: a critical re-analysis and a revised meta-analytic test of fear appeal theory. *Health psychology review*, *7*(sup1), S8–S31.

Potthoff, S., Presseau, J., Sniehotta, F. F., Johnston, M., Elovainio, M., & Avery, L. (2017). Planning to be routine: habit as a mediator of the planning-behaviour relationship in healthcare professionals. *Implementation Science*, *12*(1), 24.

Prestwich, A., Sniehotta, F. F., Whittington, C., Dombrowski, S. U., Rogers, L., & Michie, S. (2014). Does theory influence the effectiveness of health behavior interventions? meta-analysis. *Health Psychology*, *33*(5), 465.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis. *Publication bias in meta-analysis*, 1–7.

Sánchez-Meca, J., & Marín-Martínez, F. (1998). Testing continuous moderators in meta-analysis: A comparison of procedures. *British Journal of Mathematical and Statistical Psychology*, *51*(2), 311–326. doi: 10.1111/j.2044-8317.1998.tb00683.x

Schmidt, F. L., & Hunter, J. E. (2014). *Methods of meta-analysis: Correcting error and bias in research findings.* Sage publications.

Schmidt, F. L., Oh, I.-S., & Hayes, T. L. (2009). Fixed-versus random-effects models in meta-analysis: Model properties and an empirical comparison of differences in results. *British Journal of Mathematical and Statistical Psychology*, *62*(1), 97–128.

doi: 10.1348/000711007X255327

Schoemann, A. M. (2016). Using multiple group modeling to test moderators in meta-analysis. *Research synthesis methods*, *7*(4), 387–401. doi: 10.1002/jrsm.1200

Sheu, C.-F., & Suzuki, S. (2001). Meta-analysis using linear mixed models. *Behavior Research Methods, Instruments, & Computers*, *33*(2), 102–107. doi: 10.3758/BF03195354

Sidik, K., & Jonkman, J. N. (2005). Simple heterogeneity variance estimation for meta-analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, *54*(2), 367–384.

Stamatakis, E. (2006). Obesity, eating, and physical activity. Office for National Statistics/Palgrave Macmillan.

StataCorp. (2017). Stata statistical software: Release 15 [Computer software manual]. College Station, TX. Retrieved from `https://www.stata.com/`

Sterne, J. A., Bradburn, M. J., & Egger, M. (2008). Meta–analysis in stata^{TM}. *Systematic Reviews in Health Care: Meta-Analysis in Context, Second Edition*, 347–369.

Su, X., Kang, J., Liu, L., Yang, Q., Fan, J., & Levine, R. A. (2016). Smooth sigmoid surrogate (sss): An alternative to greedy search in recursive partitioning. *Computational Statistics and Data Analysis, Under Review*.

Su, X., Peña, A. T., Liu, L., & Levine, R. A. (2018). Random forests of interaction trees for estimating individualized treatment effects in randomized trials. *Stat Med*, *37*(17), 2547–2560.

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

Tanner-Smith, E. E., & Grant, S. (2018). Meta-analysis of complex interventions. *Annual review of public health*, *39*, 135–151. doi: 10.1146/annurev-publhealth-040617-014112

Team, R. C. (2017). R: A language and environment for statistical computing. *R Foundation for Statistical Computing. Vienna, Austria*.

Tessaro, I., Rye, S., Parker, L., Mangone, C., & McCrone, S. (2007). Effectiveness of a nutrition intervention with rural low-income women. *American journal of health behavior*, *31*(1), 35–43.

Therneau, T., Atkinson, B., & Ripley, B. (2017). rpart: Recursive partitioning and regression trees [Computer software manual]. Retrieved from `https://CRAN.R-project.org/package=rpart` (R package version 4.1-11)

Thompson, S. G. (1994). Systematic review: Why sources of heterogeneity in meta-analysis should be investigated. *BMJ*, *309*(6965), 1351–1355. doi: 10.1136/bmj.309.6965.1351

Thompson, S. G., & Sharp, S. J. (1999). Explaining heterogeneity in meta-analysis: a comparison of methods. *Statistics in Medicine*, *18*(20), 2693–2708. doi: 10.1002/

(SICI)1097-0258(19991030)18:20%3C2693::AID-SIM235%3E3.0.CO;2-V

Tibshirani, R., & Knight, K. (1999). Model search by bootstrap "bumping". *Journal of Computational and Graphical Statistics*, *8*(4), 671–686.

Tipton, E., Pustejovsky, J. E., & Ahmadi, H. (2018). A history of meta-regression: Technical, conceptual, and practical developments between 1974 and 2018. *Research Synthesis Methods*.

Tobacco, T. C. P. G. T., et al. (2008). A clinical practice guideline for treating tobacco use and dependence: 2008 update: a us public health service report. *American journal of preventive medicine*, *35*(2), 158–176.

Trujillano, J., Badia, M., Serviá, L., March, J., & Rodriguez-Pozo, A. (2009). Stratification of the severity of critically ill patients with classification trees. *BMC Medical Research Methodology*, *9*(1), 83. doi: 10.1186/1471-2288-9-83

van Aert, R. C., Wicherts, J. M., & van Assen, M. A. (2016). Conducting meta-analyses based on p values: Reservations and recommendations for applying p-uniform and p-curve. *Perspectives on Psychological Science*, *11*(5), 713–729.

Van Assen, M. A., van Aert, R., & Wicherts, J. M. (2015). Meta-analysis using effect size distributions of only statistically significant studies. *Psychological methods*, *20*(3), 293.

van Genugten, L., Dusseldorp, E., Webb, T. L., & van Empelen, P. (2016). Which combinations of techniques and modes of delivery in internet-based interventions effectively change health behavior? a meta-analysis. *Journal of Medical Internet Research*, *18*(6).

Van Genugten, L., Dusseldorp, E., Webb, T. L., & Van Empelen, P. (2016). Which combinations of techniques and modes of delivery in internet-based interventions effectively change health behavior? a meta-analysis. *Journal of medical Internet research*, *18*(6), e155.

Viechtbauer, W. (2005). Bias and efficiency of meta-analytic variance estimators in the random-effects model. *Journal of Educational and Behavioral Statistics*, *30*(3), 261–293.

Viechtbauer, W. (2007a). Accounting for heterogeneity via random-effects models and moderator analyses in meta-analysis. *Zeitschrift für Psychologie/Journal of Psychology*, *215*(2), 104–121. doi: 10.1027/0044-3409.215.2.104

Viechtbauer, W. (2007b). Hypothesis tests for population heterogeneity in meta-analysis. *British Journal of Mathematical and Statistical Psychology*, *60*(1), 29–60. doi: 10 .1348/000711005X64042

Viechtbauer, W. (2010). Conducting meta-analyses in r with the metafor package. *Journal of Statistical Software*, *36*(3), 1–48.

Wallace, B. C., Schmid, C. H., Lau, J., & Trikalinos, T. A. (2009). Meta-analyst: software for meta-analysis of binary, continuous and diagnostic data. *BMC medical research*

methodology, *9*(1), 80.

Wang, M. C., & Bushman, B. J. (1999). *Integrating results through meta-analytic review using sas software*. Cary, NC: SAS Institute.

Webb, T. L., Sniehotta, F. F., & Michie, S. (2010). Using theories of behaviour change to inform interventions for addictive behaviours. *Addiction*, *105*(11), 1879–1892.

Welton, N. J., Caldwell, D., Adamopoulos, E., & Vedhara, K. (2009). Mixed treatment comparison meta-analysis of complex interventions: psychological interventions in coronary heart disease. *American Journal of Epidemiology*, *169*(9), 1158–1165. doi: 10.1093/aje/kwp014

White, M., Adams, J., & Heywood, P. (n.d.). *How and why do interventions that increase health overall widen inequalities within populations?. health, inequality and society. edited by: Babones s. 2009.* Bristol: Policy Press.

Whitley, E., Batty, G. D., Hunt, K., Popham, F., & Benzeval, M. (2013). The role of health behaviours across the life course in the socioeconomic patterning of all-cause mortality: the west of scotland twenty-07 prospective cohort study. *Annals of behavioral medicine*, *47*(2), 148–157.

Williams, S. L., & French, D. P. (2011). What are the most effective intervention techniques for changing physical activity self-efficacy and physical activity behaviour—and are they the same? *Health education research*, *26*(2), 308–322.

Wilson, S. J., Polanin, J. R., & Lipsey, M. W. (2016). Fitting meta-analytic structural equation models with complex datasets. *Research synthesis methods*, *7*(2), 121–139. doi: 10.1002/jrsm.1199

Yang, Y., & Raine, A. (2009). Prefrontal structural and functional brain imaging findings in antisocial, violent, and psychopathic individuals: a meta-analysis. *Psychiatry Research: Neuroimaging*, *174*(2), 81–88. doi: 10.1016/j.pscychresns.2009.03.012

Chapter 9

Summary in Dutch (Samenvatting)

Dit proefschrift geeft de ontwikkeling weer van een nieuwe data-analyse methode op het gebied van meta-analyse. De methode integreert op een inventieve manier twee bestaande methoden, classificatie- en regressiebomen (CART) en meta-regressie, en draagt de naam meta-CART.

Hoofdstuk 1

We schetsen in dit inleidende hoofdstuk het kader van de onderzoeken in dit proefschrift en geven de motivatie weer voor het ontwikkelen van de nieuwe methode.

Meta-analyse is de analyse van gegevens van reeds uitgevoerde wetenschappelijke studies. Het voornaamste doel van een meta-analyse is het kwantitatief samenvatten van de resultaten (i.e., effectgroottes) van deze studies, liefst in één enkele maat. Om deze maat te berekenen, worden de effectgroottes van de afzonderlijke studies gemiddeld, rekening houdend met de nauwkeurigheid van de schatting van de effectgrootte per studie. Het berekenen van een dergelijke "gemiddelde" effectgrootte heeft zin als de studies een homogene groep vormen. Echter, in de praktijk is dit meestal niet het geval. We spreken dan van heterogeniteit in de effectgroottes van de studies. De uitdaging is vervolgens om deze heterogeniteit te verklaren aan de hand van karakteristieken van de studies (e.g., kenmerk van de steekproef, implementatie van de behandeling). De analyse die zich hierop richt heet moderator-analyse.

De meest gebruikelijke methode voor moderator-analyse heet meta-regressie. De predictoren in meta-regressie zijn de karakteristieken van de studies (i.e., de moderatoren) en de uitkomstmaat is de effectgrootte. Meta-regressie is minder geschikt in de volgende situaties: a) Als er veel studiekarakteristieken zijn gemeten heeft de methode niet genoeg "power" om een moderator-effect aan te tonen; b) Als onderzoekers geïnteresseerd zijn in het effect van combinaties van karakteristieken (i.e., interacties tussen moderatoren) wordt het aantal termen in de regressie-analyse al snel te veel om betrouwbare schattingen te verkrijgen.

Om bovengenoemde beperkingen van meta-regressie het hoofd te bieden, kwamen we op het idee om classificatie- en regressiebomen (CART) toe te passen in het raamwerk van meta-analyse. In de literatuur was dit nog niet eerder gedaan. Het resultaat van een CART analyse zijn subgroepen van objecten (in ons geval: studies) die meer op elkaar lijken. Dit is precies waar we naar op zoek zijn: studies die meer homogeen zijn met betrekking tot hun effectgrootte. Het idee is als eerste ruwweg uitgewerkt door Dusseldorp et al. (2014) en de voorgestelde nieuwe methode werd meta-CART genoemd. In deze eerste uitwerking werd in een eerste stap een classificatieboom gefit op de gegevens en vervolgens werd een subgroep analyse uitgevoerd (i.e., meta-regressie met één categorische predictor). De studie van Dusseldorp et al. bracht een aantal knelpunten naar boven: a) de methode werd slechts getest op één enkele bestaande dataset; b) de effectgrootte werd eerst gedichotomiseerd op de mediaan, terwijl dit waarschijnlijk informatieverlies opleverde en c) er werd geen rekening gehouden met de nauwkeurigheid van de schatting van de effectgrootte per studie.

In dit proefschrift worden deze knelpunten opgelost en wordt de methode meta-CART verder uitgewerkt in verschillende implementaties. Deze implementaties worden vergeleken en getest in uitgebreide simulatiestudies en toegepast op bestaande datasets. Daarnaast wordt er een gebruikersvriendelijk software pakket "`metacart`" voor de `R`-omgeving ontwikkeld.

Hoofdstuk 2

In dit hoofdstuk vergelijken we meta-classificatiebomen (i.e., de eerste versie van meta-CART waarbij de effectgrootte werd gedichotomiseerd) met meta-regressiebomen (i.e., zonder dichotomisatie). Daarnaast passen we gewichten toe die rekening houden met de nauwkeurigheid van de schatting van de effectgrootte per studie. Voor de gewichten ontwikkelen we drie versies. In de eerste versie worden alle gewichten gelijk gezet aan 1 (dus geen rekening houdend met de nauwkeurigheid). In de tweede versie zijn de gewichten gebaseerd op het fixed-effect meta-analysemodel en in de derde versie op het random-effects meta-analysemodel.

Om de beste grootte van de boom te bepalen maakt meta-CART gebruik van een 'pruning' regel. Door middel van kruisvalidatie wordt de goodness-of-fit (i.e., kwadratensom van de error) bepaald van bomen met verschillende groottes. De beste boomgrootte is de kleinste boom met een gekruisvalideerde fit die voldoet aan de $c*SE$-regel, waarbij c de vooraf bepaalde pruning-parameter is en SE de standaardfout is van de laagste gekruisvalideerde kwadratensom van de error. In CART wordt meestal ge-werkt met $c = 1$. In de huidige studie onderzoeken we wat in meta-CART de beste waarde is voor de pruning parameter. We onderzoeken de volgende drie waardes: 0, 0.5, of 1. Dit alles resulteert in een 3 (waarde van c) bij 3 (type gewicht) bij 2 (type boom) ontwerp.

In een uitgebreide Monte-Carlo simulatiestudie testen we de resultaten van deze meta-CART ontwerpen. We creëren artificiële data aan de hand van verschillende ware modellen (scenario's). Deze scenario's verschillen in de complexiteit van het moderator-effect. Ook is er een scenario waarbij de studies een homogene groep vormen en er geen moderator effect aanwezig is. In dit laatst genoemde scenario treedt er een Type I fout op als meta-CART concludeert dat er wel sprake is van een moderator. Voor de overige ware scenario's, waarin één of meerdere moderator-effecten aanwezig zijn, bepalen we de power en in hoeverre meta-CART de ware structuur kan vinden (m.b.t. het aantal subgroepen en de gekozen moderatoren).

De conclusie van de simulatiestudie is dat een meta-regressieboom met random-effects gewichten en een pruning parameter van 0.5 de beste resultaten oplevert. Zelfs met een streng pruning parameter ($c = 1$) blijkt de Type I fout van meta-classificatiebomen niet onder controle te krijgen en er wordt besloten om in het vervolg alleen meta-regressiebomen te gebruiken.

Hoofdstuk 3

In Hoofdstuk 3 stellen we twee nieuwe algoritmes voor die het algoritme van CART combineren met subgroep analyse: één voor het fixed-effect model en één voor het random-effects model. Hierdoor is de twee-stappen-benadering van meta-CART overbodig geworden.

In het eerste algoritme wordt het fixed-effect model geïntegreerd in het partitioneringsalgoritme. Op elegante wijze wordt aangetoond dat het toepassen van fixed-effect gewichten gelijk is aan het minimaliseren van de som van de Q-statistiek in de kindknopen. Deze knopen ontstaan na een split van de ouderknoop op een drempelwaarde van een moderator, waardoor de groep van studies in de ouderknoop wordt verdeeld in twee kindknopen. Een knoop representeert dus een subgroep van studies. De binnengroep Q-statistiek is een maat voor de heterogeniteit van de effectgroottes van de studies die tot de knoop behoren. De studie laat zien dat door het minimimaliseren van de binnengroepen Q-statistiek (na iedere split) de tussengroepen Q-statistiek wordt gemaximaliseerd. Deze tussengroepen Q-statistiek is een maat voor de sterkte van het moderator-effect. Met behulp van een χ^2 test wordt getoetst of de tussengroepen Q-statistiek van de boom na pruning significant is.

In het tweede algoritme wordt het random-effects model geïntegreerd in het partitioneringsalgoritme. In de stapsgewijze procedure (zoals toegepast in Hoofdstuk 2) wordt de residuele variantie (σ^2_τ) aan het begin van de eerste stap (de boomconstructie) berekend voor het berekenen van de random-effects gewichten en σ^2_τ wordt opnieuw berekend aan het begin van de tweede stap voor de subgroep analyse. Echter, het vastzetten van σ^2_τ in de random-effects gewichten van de eerste stap is niet consistent omdat in feite σ^2_τ

verandert na iedere split van het partitioneringsalgoritme. In het nieuwe algoritme wordt hiermee rekening gehouden en wordt de residuele variantie (van de tot dan toe gefitte boom) opnieuw geschat na iedere split. Er wordt aangetoond dat in dit nieuwe algoritme in iedere split de random-effects Q-statistiek wordt gemaximaliseerd.

De uitgebreide simulatiestudie toont aan dat indien de juiste pruning regel wordt gebruikt en het aantal studies in de steekproef groot genoeg is, de nieuwe meta-CART zeer goede resultaten laat zien (power > 0.80 en Type I fout < 0.05). Het minimum benodigde aantal studies varieert tussen de 40 en 120, afhankelijk van de complexiteit en de sterkte van de ware mderator-effecten. Ook de steekproefgrootte binnen de studies, het type moderator (dichotoom, nominaal, of continu) en de grootte van de residuele variantie spelen een rol.

Hoofdstuk 4

In dit hoofdstuk wordt nader ingegaan op het softwarepakket `metacart`, versie 2.0, voor de `R`-omgeving. In deze versie zijn de nieuwe algoritmes van hoofdstuk 3 geïmplementeerd. Bovendien wordt een nieuw aspect geïntroduceerd in het algoritme van de random-effects meta-regressiebomen: de "look-ahead" strategie.

Bij het bepalen van de eerste beste split (i.e., moderator en splitpunt) wordt in de "look-ahead" strategie twee stappen vooruit gekeken. Dit betekent dat het partitiecriterium (de tussengroepen Q-statistiek) berekend wordt na twee splits voor iedere combinatie van moderatoren en voor ieder mogelijke split op ieder van deze moderatoren. De combinatie die de hoogste tussengroepen Q-statistiek oplevert wordt gekozen als de beste. Dit resulteert in twee splits tegelijk en dus in een boom met drie eindknopen (subgroepen). Deze drie eindknopen functioneren vervolgens als mogelijke "ouder" voor de volgende split, waarbij de gewone "greedy" zoekstrategie wordt gebruikt (geen look-ahead). Het voordeel van de look-ahead strategie om de eerste twee splits te bepalen is dat de boom niet gedomineerd wordt door een sterk hoofdeffect van een moderator. In plaats daarvan zoekt de look-ahead naar de sterkste interactie.

In het hoofdstuk worden de twee hoofdfuncties van `metacart` uitgelegd: de functie die een fixed-effect meta-regressieboom fit, `FEmrt`(), en degene die een random-effects meta-regressieboom fit, `REmrt`(). De functie `FEmrt`() maakt gebruik van het bestaande `R`-package `rpart`. De functie `REmrt`() is geheel nieuw. De input argumenten van beide functies verschillen iets van elkaar en worden stuk voor stuk uitgelegd. De functies resulteren in een S3-object met bijbehorende "print", "plot", "summary" en "predict" functies.

Aan de hand van drie voorbeelden wordt `metacart` geïllustreerd. In het eerste voorbeeld met de dataset "dat.bourassa1996" wordt meta-CART met meta-regressie vergeleken. De effectgrootte van de studies is een log odds ratio. Een hogere waarde geeft een sterker verband weer tussen links of rechts handig zijn en linker of rechter oogdominantie. Meta-

regressie wordt toegepast om de vraag "welke moderatoren hebben een invloed op de hand-oog associatie" te beantwoorden. Vervolgens wordt met meta-CART de vraag beantwoord "welke combinaties van (categorieën) van moderatoren hebben een invloed op de hand-oog associatie?". De oplossing van meta-CART laat in een oogwenk zien welke categorieën van een invloedrijke multinominale moderator samengevoegd kunnen worden. Het resultaat blijkt goed overeen te stemmen met de hypotheses in het artikel waarin de data oorspronkelijk werden gepresenteerd.

In het tweede voorbeeld wordt `metacart` toegepast op een subset van de dataset van Michie et al. (2009) met studies die het effect onderzoeken van een gezondheidsbevorderende interventie (beweeg- of voedingsinterventie). De subset bevat vijf dichotome moderatoren die aangeven of een bepaalde gedragsveranderingstechniek wel of niet is gebruikt in de interventie. De effectgrootte is Hedges' g, het gestandaardiseerde gemiddelde verschil op de uitkomstmaat tussen de interventie- en controlegroep. De onderzoeksvraag is "Welke combinaties van gedragsveranderingstechnieken zorgt voor het grootste interventie-effect?". Zowel een fixed-effect meta-regressieboom analyse als een random-effects meta-regressieboom analyse worden toegepast op de data. Alleen de fixed-effect meta-regressieboom analyse vindt moderator effecten: een interactie tussen twee gedragsveranderingstechnieken (T1: verschaft informatie over het verband tussen gedrag en gezondheid en T4: stimuleert intentievorming). De studies die beide technieken toepassen resulteren in het grootste gemiddelde effect van de interventie.

In het derde voorbeeld wordt het voordeel van de look-ahead strategie geïllustreerd aan de hand van een gesimuleerde dataset. Het ware model onder de data is gebaseerd op Tibshirani en Knight (1999). DIt model bevat een disordinale interactie tussen twee moderatoren die erg lastig te ontdekken is zonder look-ahead strategie.

De conclusie is dat de kracht van meta-CART ligt in het detecteren van combinaties van moderatoren en combinaties van categorieën van moderatoren die zo goed mogelijk de heterogeniteit in de effectgroottes kunnen verklaren. Het `R`-package `metacart` bevat functies die voor fixed-effect meta-analyse en random-effects meta-analyse geschikt zijn, met opties die de gebruiker gemakkelijk kan aanpassen.

Hoofdstuk 5

De systematische review en meta-analyse die beschreven is in dit hoofdstuk is uitgevoerd in samenwerking met onderzoekers uit het Verenigd Koninkrijk en Canada. In de systematische review zijn gerandomiseerde effect studies (RCTs) opgenomen waarin personen aan een gezondheidsbevorderende interventie (gezonde voeding, fysieke activiteit, of stoppen-met-roken interventie) werden toegewezen of aan een controle groep. De doelgroep van de interventies zijn volwassenen met een laag inkomen. In totaal zijn 45 interventies opgenomen, beschreven in 35 artikelen.

In de meta-analyse wordt exploratief onderzocht welke (combinaties van) inhoudelijke componenten en manieren-van-aanbieden van de interventies geassocieerd zijn met de effectgrootte. De gebruikte effectgrootte is Hedges' g voor de gedragsverandering in voeding of bewegen. Hedges' g is het gestandaardiseerde verschil in gemiddelde verandering in gezondheidsgedrag tussen de interventie- en de controlegroep, waarbij een correctiefactor wordt toegepast voor kleine steekproeven. Een positieve Hedges' g betekent dat de interventiegroep het beter doet dan de controlegroep. Voor stoppen-met-roken is de gebruikte effect grootte het relatieve risico op stoppen met roken (i.e., abstinentie).

In univariate random-effects meta-regressies (voor continue interventiekenmerken) en subgroep analyses (voor categorische kenmerken) wordt het moderator effect geschat van in totaal 46 verschillende gedragsveranderingstechnieken en 14 manieren-van-aanbieden. De analyses worden uitgevoerd per type interventie (gezonde voeding, fysieke activiteit, of roken) en alleen de moderatoren die door minstens drie interventies van dat type zijn gebruikt worden onderzocht.

De meta-CART analyses richten zich op het detecteren van de meest invloedrijke combinaties van moderatoren per type interventie. Omdat de steekproef van studies niet zo groot is, worden in de meta-CART analyses alleen de moderatoren onderzocht die in de univariate analyses een significant effect lieten zien. Zowel fixed-effect als random-effects meta-regressiebomen worden gefit per type interventie. Hiervoor is gekozen omdat beide voordelen hebben: fixed-effect meta-regressiebomen hebben een grotere power (zijn liberaler), terwijl random-effects meta-regressiebomen de controle van de Type I fout maximaliseren (zijn conservatiever). Bij de fixed-effect meta-CART analyse is er een sterkere aanname: de heterogeniteit in de effectgroottes van de studies wordt volledig verklaard door de verschillende subgroepen (de moderatoren) en de binnenknopen steekproeffout, er is geen residuele variantie. Dit brengt het totaal op zes meta-CART analyses. In deze samenvatting gaan we in op de resultaten hiervan.

Voor de gezonde voedingsinterventies geeft de random-effects meta-CART analyse als resultaat een boom met twee splits en de tussengroepen Q-statistiek is significant. Dit suggereert dat de interactie van de twee moderatoren (gerepresenteerd door de boom met twee splits) de heterogeniteit in de effect groottes van de studies kan verklaren. Het blijkt dat interventies die geen "feedback op gedrag" geven gemiddeld de hoogste effectgrootte hebben ($\bar{g} = 0.36$). Van de groep interventies die wel "feedback op gedrag" geven resulteren de face-to-face interventies in een gemiddeld hogere effect-grootte ($\bar{g} = 0.23$) dan de interventies die niet face-to-face worden aangeboden ($\bar{g} = 0.10$).

De fixed-effect meta-CART analyse resulteert in een meta-regressieboom met dezelfde twee splits als de random-effects meta-regressieboom en voegt daaraan een split toe van de groep interventies die "geen feedback geven". De split is op de moderator "zelf-monitoren van gedrag". De interventies die geen feedback geven, maar wel het zelf-monitoren van gedrag aanbieden, resulteren in een hogere gemiddelde effectgrootte ($\bar{g} = 0.48$) dan de

interventies die geen feedback geven en geen zelf-monitoren aanbieden ($\bar{g} = 0.31$).

Voor de fysieke activiteitsinterventies resulteert de random-effects meta-CART analyse in een meta-regressieboom met één knoop (i.e., de "root node"). Dit betekent dat er geen moderator effect gevonden is. De fixed-effect meta-CART analyse resulteert in een meta-regressieboom met twee splits, dus drie eindknopen. Interventies die aangeboden worden in gezondheidscentra resulteren in de laagste gemiddelde effectgrootte ($\bar{g} = -0.002$). Interventies die aangeboden worden in gemeenschapsruimtes (bijv. het park) of thuis maar geen instructies geven over hoe een gedrag uitgevoerd moet worden resulteren in een lagere gemiddelde effect grootte ($\bar{g} = 0.21$) dan interventies die in dezelfde omgeving worden aangeboden en wel instructies geven over de uitvoering van een gedrag ($\bar{g} = 0.42$).

Voor de stoppen-met-roken interventies wordt er door zowel de fixed-effect als de random-effects meta-CART analyse geen moderator effect gevonden.

De conclusie is dat bij volwassenen met een lager inkomen het wel of juist niet includeren van bepaalde gedragsveranderingstechnieken en de omgeving (waar een interventie wordt aangeboden) een grote invloed kunnen hebben op het effect van interventies die gericht zijn op het bevorderen van gezonde voeding en fysieke activiteit. Voor stoppen-met-roken interventies kunnen we geen uitspraak doen over invloedrijke interventiekenmerken.

Hoofdstuk 6

De methode meta-CART heeft een tweetal minpunten die ook voorkomen bij sommige andere partioneringsalgoritmen. Ten eerste maakt het algoritme gebruik van een "greedy search" strategie, waarbij alle mogelijke moderatoren en splitpunten afgegaan worden om tot een split te komen. Dit betekent dat meta-CART niet kan garanderen dat de gevonden oplossing globaal optimaal is en de resultaten kunnen instabiel zijn. Ten tweede is statistische inferentie in de gedetecteerde subgroepen (i.e., de eindknopen van de boom) lastig omdat deze groepen niet vooraf gespecificeerd zijn, maar zijn verkregen door het zoeken in de data.

Om aan deze minpunten het hoofd te bieden, stellen we in dit hoofdstuk de volgende vier uitbreidingen van meta-CART voor:

i Voor de boomconstructie wordt de smooth sigmoid surrogate (SSS) methode gebruikt; hierdoor verwachten we dat lokale optima minder vaak voorkomen;

ii Voor de boomconstructie wordt de look-ahead strategie die beschreven is in hoofdstuk 4 ook uitgewerkt voor fixed-effect meta-regressiebomen; dit zorgt voor een verlaging van de kans op lokale optima;

iii Voor de statistische inferentie in de subgroepen wordt een permutatie-test gebruikt om de significantie van de tussenknopen Q-statistiek te bepalen. Deze test wordt

gebruikt om te bepalen of er sprake is van een moderatoreffect of niet.

iv Om in de subgroepen te corrigeren voor het overoptimisme in de schattingen van het betrouwbaarheidsinterval van de gemiddelde effectgrootte is een nieuwe bootstrap procedure ontwikkeld, die rekening houdt met de nauwkeurigheid van de effectgroottes.

De resultaten van een uitgebreide simulatiestudie laten zien dat de bovengenoemde uitbreidingen van meta-CART veelbelovend zijn. Het toepassen van de SSS-methode en de look-ahead strategie verlagen de kans op fout-positieve resultaten (oftwel het vinden van spurieuze moderatoreffecten) en verhogen de kans op het detecteren van de juiste moderatoreffecten. De permutatietest verlaagt de kans op fout-positieve resultaten en de bootstrap procedure geeft een juiste correctie van de schatting van de residuele variantie.

De conclusie van dit hoofdstuk is dat de vier voorgestelde uitbreidingen zorgen voor een verbeterde prestatie van meta-CART. De uitbreidingen zullen worden geïmplementeerd in `metacart`, versie 3.0.

Epiloog

In een meta-analyse is vaak sprake van heterogeniteit in de gevonden effectgroottes van de studies. Meta-CART is een nuttige techniek om schijnbaar tegenstrijdige bevindingen van studies te verklaren. Door invloedrijke studiekenmerken (i.e., moderatoren) op te sporen, partitioneert meta-CART de studies in subgroepen die meer homogeen zijn met betrekking tot hun effectgroottes. In dit proefschrift is meta-CART verder ontwikkeld met als doel het zo goed mogelijk detecteren van homogene subgroepen in meta-analyse door gebruik te maken van (combinaties van) moderatoren.

Door middel van simulatiestudies hebben we de prestaties van meta-CART getest. In de simulatiestudies van hoofdstukken 2 en 3 zijn ware scenario's gebruikt die door een boom werden gerepresenteerd. Daarmee was het relatief gemakkelijk om de ware structuur met meta-CART terug te vinden. In hoofdstuk 6 bevatten de ware scenario's ook andere, meer complexe modellen. Ook in deze modellen slaagt meta-CART erin om (gedeeltes) van de ware moderatoren te ontdekken zonder spurieuze effecten te vinden.

Op basis van de resultaten van de simulatiestudies zijn er richtlijnen opgesteld voor het gebruik van meta-CART. Over het algemeen zijn random-effects meta-regressiebomen te verkiezen boven fixed-effect meta-regressiebomen omdat random-effects meta-regressiebomen rekening houden met de variantie die niet verklaard wordt door de moderatoren. Daarnaast wordt aangeraden de vier uitbreidingen van hoofdstuk 6 te gebruiken. Voor het bepalen van de juiste boomgrootte raden we de $0.5 * SE$ pruning-regel aan. Deze regel geeft een goede balans tussen de Type I fout en de power. Om meta-CART toe te passen met voldoende power ($\geq .80$) zijn de volgende randvoorwaarden geformuleerd:

- een minimale grootte van 40 studies voor het detecteren van een hoofdeffect van een moderator en een tweeweg-interactie tussen moderatoren (i.e., een boom met twee of drie eindknopen);

- een minimale grootte van 80 studies voor het detecteren van meer complexe interacties (i.e., een boom met vier of meer eindknopen).

Naast artificiële datasets is meta-CART in dit proefschrift toegepast op echte datasets. Deze datasets verschilden in de gemeten effectgrootte, bijvoorbeeld het gestandaardiseerde gemiddelde verschil (Hedges' g) en de log odds ratio. Meta-CART kan toegepast worden op ieder type effectgrootte die voldoet aan de centrale limietstelling. Als de verdeling van de effectgrootte erg scheef is, kan een transformatie toegepast worden (bijv. Fisher's Z transformatie).

Er zijn een aantal praktische moeilijkheden bij het toepassen van meta-CART. Ten eerste weet de onderzoeker meestal niet van te voren hoeveel studies er uiteindelijk gevonden worden die geschikt zijn voor de meta-analyse en of er sprake is van complexe moderatoreffecten in de data. Daardoor is het lastig om te bepalen of de steekproef groot genoeg is voor een meta-CART analyse (uitgaande van bovengenoemde randvoorwaarden). Dit maakt het lastig om een meta-CART analyse te pre-registreren (i.e., het vooraf registreren van de studie in het kader van "open science"). Ten tweede bevatten de studiekenmerken (de moderatoren) vaak missende waarden omdat niet alle studies de kenmerken rapporteren. Een strategie kan zijn om in een onderzoeksveld een lijst op te stellen met belangrijke moderatoren die studies dienen te rapporteren (bijv. in het veld van de gezondheidspsychologie is de taxonomie van gedragsveranderingstechnieken opgesteld). Ten derde wordt in meta-analyses vaak gewaarschuwd voor publicatie bias. Dit fenomeen kan ook de resultaten van een meta-CART analyse beïnvloeden. Hoe hiervoor te corrigeren in meta-CART is een uitdaging voor toekomstig onderzoek.

Een andere richting voor toekomstig onderzoek is het toepassen van de vier voorgestelde uitbreidingen in hoofdstuk 6 op andere boomtechnieken, bijv. in het raamwerk van meta-analyses van individuele patiëntgegevens.

De conclusie van dit proefschrift is dat meta-CART een veelbelovende techniek is om subgroepen van studies op te sporen die meer homogeen zijn wat betreft hun effectgrootte. De techniek kan worden toegepast op meta-analyse gegevens met meerdere moderatoren die van verschillende aard mogen zijn (dichotoom, nominaal, ordinaal en numeriek). Ik hoop dat mijn werk de interesse van andere onderzoekers aanwakkert om bomen in meta-analyse te gaan gebruiken.

Chapter 10

Curriculum Vitae

Xinru Li was born in Baiyin (in English it means 'silver'), China, on March 29, 1988. In 2007, she obtained a Bachelor degree in Life Science at Fudan University, China. From 2007 to 2012, she worked as a PhD student on Molecular Genetics, in School of Life Science at Fudan University. During that period, she realizes that her major research interest was not experimental biology, but the analysis of data collected from experiments. Therefore, she quit the PhD program on Molecular Genetics in 2012, and came to Leiden University, Netherlands, to pursuit a Master degree in statistics. In 2014, she graduated cum laude in Statistical Science for the Life and Behavioral Sciences. After that, she worked as a PhD student in applied statistics in the Mathematical Institute at Leiden University from 2014 to 2019. This dissertation is the result of the project 'meta-CART', on which she had been working during this period. Her research interests are: tree-based models, prediction, classification, interaction effects and meta-analysis. Since 2019, she has a position as a Data Scientist at Google, Switzerland.

Chapter 11

Acknowledgment

During the process of writing this monograph, I was inspired and supported by many people. First of all, I would like to express my sincere gratitude to my supervisors Elise Dusseldorp and Jacqueline Meulman for their encouragement, advice, and inspiring discussions with herb tea or red wine. I have learned a lot from them, and this thesis would not have been possible without their patient guidance. In particular, I want to thank Elise for our awesome 'dishes ritual' (i.e., washing dishes together before we have tea), which gave me many ideas and thoughts for this thesis. Also, I want to thank all my colleagues at Leiden University for creating such a pleasant work atmosphere.

I am grateful to Xiaogang Su for his involvement with my work and for hosting me at UTEP. I had a great time in El Paso and learned a lot from him. The idea of using smooth sigmoid surrogate in meta-CART was inspired by his previous work on random forests of interaction trees. When working on Chapters 4, 5 and 6, I received lots of help from him for improvements on both the methodology and the efficiency of the software.

I want to thank Prof. Ingram Olkin for encouraging us to submit Chapter 3 to Research Synthesis Methods.

I want to thank Robbie van Aert for his inspiring suggestions on Chapters 3 and 6.

I want to thank Marcel van Assen for reading my thesis and giving lots of comments on Chapter 6.

I thank all my friends, both in the Netherlands and in China, for their support (e.g., kindly pretending to be interested when I was talking about my work).

I want to thank my beloved family for their contribution both to my personal growth and to my work. I thank Kaihua for his help in all sorts of ways, especially those heated debates over whose way of programming is smarter. I thank my mother for her unconditional love and support.