**Performance of meta-analysis heterogeneity detection techniques when moderator effects are non-normally distributed: A simulation study**

**Abstract**

Meta-analysis has a major prestige and influence in contemporary science. Nevertheless, a serious concern is heterogeneity that is included between the studies in meta-analysis, which is caused due to clinical and/or methodological diversity. Heterogeneity causes a challenge to aggregate data, but it also offers a way to find moderators that could have impact on the found effect size. If the causes of heterogeneity are known beforehand, they can be explained using meta-regression, but if this is not the case, there is need for an exploratory approach. Various methods have been developed to calculate their contribution. Most common is the classic method random-effects meta-analysis. A recently developed method departed from the linear models, and introduced metaCART, which uses single trees to explore heterogeneity. Which was a step in the right direction, but in response to overcome the limitations of single trees by using random-forests, MetaForest was developed. Which demonstrated superior predictive performances over all other approaches when moderator effects were normally distributed. For a more representative conclusion, this study will compare these techniques when moderator effects are non-normally distributed. By the hand of a simulation we found out that random-effects meta-analysis performed superior and metaCART inferior. A fundamental comment for this unexpected result is the ‘home advantage’ effect. We discuss how to prevent such effects in future research and give a possible insight in which situations linear models or random-forests should be used to explore heterogeneity in meta-analysis.

**Introduction**

Meta-analysis has become a popular and top of tiers approach, and has major prestige and influence in contemporary science (Tatsioni & Loannidis, 2008).The inventor of the label ‘Meta-analysis’, an education researcher, defined meta-analyses as follows: “the statistical analysis of a large collection of analysis results from individual studies for the purpose of integrating the findings (Glass, 1976).” It is basically a study about studies, used to get a unified result. A researcher reviews already published articles on a certain topic, and analyses the results to find general trends across those studies. Combining data may improve statistical power when there are multiple studies on an explicit research question, but those studies can be underpowered or are not designed to give an answer to that research question. And therefore, meta-analysis can be very misleading and lead to wrong decision-making in for example healthcare or other scientific fields (Loannidis, 2016).

The main concern is the statistical heterogeneity between the studies included in the meta-analysis. Statistical heterogeneity in meta-analysis means that the included studies their populations, samples or results are different (Gravetter & Wallnau, 2016). Which can be caused due to clinical and/or methodological diversity. Clinical heterogeneity is about variability that originates because of distinct populations, outcomes, interventions or follow-up times. Methodological heterogeneity relates to different designs and quality (Kontopantelis & Reeves, 2010). Further in this study we will refer to statistical heterogeneity simply by heterogeneity. The opposite of heterogeneity is homogeneity; in case of a meta-analysis it means that the study data included in the meta-analysis are identical (Gravetter & Wallnau, 2016).

Heterogeneity causes a challenge to aggregate data, but it also offers a way to found out which discrepancies between studies have an impact on the found effect size, these discrepancies are known as ‘moderators’ (van Lissa, 2017). There are two classic approaches to meta-analysis to account for the influence of those moderators. The first one is fixed-effects meta-analysis. That assumes each observed effect size is an estimation of the underlying true effect size, and is subject to sample error (Hedges & Vevea, 1998). Therefore, for an amount of *k* studies, the observed effect size of each individual study (for ) is given by (van Lissa, 2017):

where ) (1)

Fixed-effects meta-analysis considers sampling error to be the only cause of variance that influences the observed effect size. Studies with a large sample size will, as a result of this, produce more precise estimations of the underlying true effect size (Schmidt, Oh & Hayes, 2009). Therefore, large sample sizes will contribute more to the weighted mean than small sample sizes. With that said, fixed-effects weights are defined by van Lissa (2017): “as the reciprocal of the effect size variances”:

(2)

The weighted average can only be considered an estimate of the true effect size if the assumption can be made that each study has the same underlying true effect. If this assumption is not the case, the weighted average is pondered as a summary of the observed effect sizes (Hedges & Vevea, 1998; Borenstein, Hedges, Higgins & Rothstein, 2010).

Thus, for an accurate estimate of the fixed-effects meta-analysis model, it starts with the assumption that the true effect is even in al studies. However, this assumption is implausible in plenty of meta-analysis. Especially in social sciences, the behaviour of people is extremely diverse and the contextual conditions for all humans vary to a great degree (Aronson, Wilson & Akert, 2016). A highly likely consequence is that this will lead to a huge amount of possible moderators (Caserio, 2014). Also studies that examine same or similar research questions often differ. Caused by differences in for example cultures of research populations and used methods or instruments (Neuman, 2011). Even in replication studies there are sometimes moderators that are unanticipated (Kunert, 2016). This leads very often to an eventual poor performance of the fixed-effects meta-analysis model (Snijders, 2005).

The second classic meta-analysis approach is random-effects meta-analysis. Instead of assuming that each observed effect size is an estimation of the underlying true effect size this approach considers the true effect size to follow a distribution. In other words, that each study estimates a study-specific true effect. Also, the differences between observed effect sizes originate from the sampling error and the between-study variance (Borenstein et al., 2010). The formula of the random-effects meta-analysis model is (van Lissa, 2017):

where (3)

where

In the random-effects model, the summary effect is the same as in the fixed-effects model, a weighted mean of the observed effect sizes. Consequently, the weights are assigned to individual studies that are based on the sources of variance assumed to influence the observed effect size. Instead of assuming that only sampling error influences the effect size, random-effects also assumes that the underlying true effect sizes follow a distribution. Therefore, all studies contribute info about the distribution (Borenstein, Hedges & Rothstein, 2007; Chen, Manning & Dupuis, 2012). About the weighting, they apply this through debilitation of study weights relative to the estimated amount of between-study variance. In this manner, when observed effect sizes are more heterogeneous, the more equally each study contributes to the weighted mean. Eventually, the summary effect is the point estimate of the true effect size distributions mean (van Lissa, 2017).

(4)

Most simulation studies that compared both classic meta-analytical models concluded that the random-effects model obtained a more precise estimate, and the confidence interval had a more (close to) correct coverage in comparison with the fixed-effects model. Fixed-effects meta-analysis is only an option if the studies in the meta-analysis are estimating the same common effect. When there will be between-study heterogeneity in true effects, a random-effects model is ought to be adopted (Higgins, Thompson & Spiegelhalter, 2009).

Regardless of the capability of the random-effects model delivering precise estimates in meta-analysis. Published meta-analysis barely explains many moderators, sometimes none at all. There are a couple of reasons to explain this: Riley, Higgins & Deeks (2011) concluded that the number of studies is oftentimes overly small. As a result of this heterogeneity cannot be inspected accurately. Also, there is an insufficiency in regard to techniques to downgrade the amount of possible moderators to a feasible number (Thompson & Higgins, 2002). This results in a large amount of moderators to be examined and a low quantity of studies. These kinds of conditions do not have a solid fit in the classic meta-analysis approaches (van Lissa, 2017). For this obstacle there is a variable selection technique needed, that identifies moderators as strong or weak influencers of the observed effect size.

An alternative that can perform variable selection, are tree-based models. Besides this fact, these kinds of models have numerous other advantages over linear models. Tree-based models: can be used for any data type, are easy to represent visually, require little data preparation and got larger power than linear regressions when moderators exceed observations in quantity. They are also more flexible in handling moderator interactions and non-linearity. As a result of that, they are better in modelling the complicated nature of human behaviour (Earp & Trafimow, 2015). Decision trees split from the top down and group data in so-called ‘sub-nodes’, in which the data’s aspects are most homogeneous. The goal is to split to get the sub-nodes as uniform as possible, which can be until fully homogenous groups, or if a pre-specified touchstone is reached.

Li, Dusseldorp and Meulman (2017) introduced metaCART, an approach that applies classification and regression trees (CART), in terms of single trees. It identifies interactions, and thereafter it utilizes subgroup meta-analysis to test the significance of the moderator effects. Recent studies investigated the efficiency of metaCART, which indicated it to be a stable technique to explore moderator effects on the observed effect size. Still, singletree based models have some limitations and for that reason, metaCART as well. First of all, tree models are unstable, small fluctuations that are utilized to make the model have a possibility to lead to considerable alterations in the constructions of the tree (Dwyer and Holte, 2007). Second, it has problems with seizing linearity, because it only makes ‘twofold splits’ (Steyerberg, 2019). At last, tree-based models are susceptible to overfitting (Hastie et al, 2009).

There are also more complex tree-based models, known as random-forests, which surmount most of the disadvantages of singletree. This variant incorporates multiple decision trees, and combines results from those trees to create a single model with a more accurate estimate (Breiman, 2001). The essential idea behind is know as the ‘wisdom of crowds’, a large number of relatively uncorrelated trees operating as a group will outperform any of the individual elements. The somewhat low correlation between the models is fundamental, because uncorrelated models are able to produce ensemble predictions with a higher accuracy that any individual prediction. This is because the trees preserve each other from their own singular errors (Genuer, Poggi & Tuleau-Malot, 2010). The lower tendency to overfitting is another advantage of random forests over single trees (Bühlmann & Yu, 2002). As well as the possibility to predict cases that are not components of the bootstrap sample of the tree. This kind of measure is known as out-of-bag error, which is an approximation of the cross-validation error, and provides proper estimates of the prediction accuracy in further samples (Hastie et al., 2009).

An alternative to explore heterogeneity in meta-analysis with a singletree-based method is MetaForest. A technique developed by van Lissa (2017), designed to overcome the lacking’s of singletrees by using random forests. MetaForest applies random-effects or fixed-effects weights to random forests. Based on two simulation studies, van Lissa (2017) examined the performance of fixed-effects, random-effects and unweighted MetaForest and compared it to metaCART. All three variants of MetaForest had in both studies a superior predictive performance to metaCART, consequently random-effects MetaForest outperformed fixed-effects and unweighted MetaForest. The study displayed also other advantages from random forests over singletrees. It had greater power, was able to make better predictions, gave estimates of the cross-validation error and yielded useful measures of variable importance and partial prediction plots (van Lissa, 2017). MetaForest can at the moment be considered as the best working technique to explore heterogeneity in meta-analysis.

However, there is an important feature to prove before we can make such an assumption. The underlying data generating models in the two simulation studies of van Lissa (2017) only included normal distributed moderators. Renouncing from normal distributions may affect the performance of the model, but since normal distribution in real-life data is more an exception than a normal state of affairs (Micceri, 1989), it is entirely possible that procedures are affected by skewness, leverage, balance etc. It is important to know how MetaForest performs in these kinds of situations. Despite this, it is hard to find studies that have researched the effect of non-normal study effects on the performance of meta-analysis methods. Studies about this topic have mostly been shortened to delimited comparisons of distribution free methods to commonly used parametric methods, which were than under the specific condition that included normality. Only Kontopantelis & Reeves (2010) compared eight different frequently used meta-analysis techniques, they found some differences between those techniques, but the results were remarkably steady across divergent effect size distributions, that ranged from normal through the utmost skewness and kurtosis. This is an important finding that can indicate a possible robustness from MetaForest against even very severe violations of the assumptions of normality.

Another factor that plays a role in affecting a model’s performance, even if it is normally distributed, is the amount of studies that is included in the meta-analysis. Simulations were the moderators were normal shown that the coverage levels of every meta-analysis methods get affected when the amount of studies are small, even in modest degrees of heterogeneity (Brockwell & Gorden, 2001; Kontopantelis & Reeves, 2010). It is common for meta-analysis methods to require a large number of studies in order to estimate the residual variance τ2 with a decent precision (Jackson & White, 2018). Yet, in van Lissa (2017), that only presented estimates of τ2 based on the raw data, we saw that MetaForest had certain robustness against a low number of studies. If moderators were continuously distributed, MetaForest had sufficient power at approximately 20 studies. Therefore, we expect MetaForest to have sufficient power when there are non-normal distributed moderators active as well.

A third, and obvious factor, that is classified as relevant to model performance is heterogeneity among studies being meta-analysed (Kontopantelis & Reeves, 2011; Jackson & White, 2018). Coverage from models degrades when the residual heterogeneityincreases, mostly when the amount of studies is small (Brockwell & Gorden, 2001). Considering that all models their performance is linked to the accuracy of the estimate. According to Sidik & Jonkman (2007), it is generally the case that the larger true between-study variance is, the more biased the estimate can be, which diminishes the performance of the method. In addition, the residual heterogeneity estimations are specified to use population-averaged within study variances, but in reality, study-specific variances are used, what can lead to bias (Bohning et al., 2002).

With the use of a simulated dataset we will compare the performances of: random-effects, fixed-effects and unweighted MetaForest, metaCART and random-effects meta-analysis, when the moderator effects that are included in the meta-analysis are non-normally distributed. Simulations studies are a meaningful instrument in determining the robustness of statistical approaches (Auerswald & Moshagen, 2015). They are used to achieve empirical results about the performance of statistical methods in particular scenarios. In contrast to the more general analytic results, which can cover many scenarios. This can make it impossible, or very hard, to obtain analytic results. Simulation studies come into their own if methods form the wrong assumptions, or in case of messy data they can assess flexibility of methods in such kind of situations. This is not always the case with analytic results that may assume that data arises from a specific model (Morris, White & Crowther, 2019). Thus, simulations are useful for evaluation of new methods like MetaForest and for the comparison with alternative methods like metaCART and the classic approaches. These simulations are computer experiments that create data by means of pseudorandom sampling, which come from known probability distributions. The main reason behind simulation studies is to be able to understand the behaviour of certain statistical methods. There is always some true data known from the generation process what allows us to consider properties of methods, like bias (Morris, White & Crowther, 2019).

**Simulation study**

This simulation is, to a certain extent, a replication of simulation 2 carried out by van Lissa (2017). We have the same performance criteria, comparable design factors, and there are the same weighted versions of MetaForest included: random-effects MetaForest, fixed-effects MetaForest and unweighted MetaForest. Differences are: the simulation contains the last CRAN-version of metaCART published by Li et al., (2019), the addition of the classic model random-effects meta-analysis. Furthermore, we included moderators that are continuous and non-normally distributed instead of normal distributed moderators.

**Performance criteria**

In this study we evaluated three performance criteria: 1) The algorithms predictive performance, 2) the algorithms power, and 3) their ability to perform variable selection.

Since meta-analysis is typically formalized as a regression problem. We defined the predictive performance for every algorithm as prediction accuracy, operationalized as the R-squared *R2*.A statistical measure that represents the proportion of the variance from a dependent variable that is explained by an independent variable or multiple variables in a statistical model (Gravetter & Wallnau, 2016).

(5)

Power is defined as the algorithms ability to detect the existence of moderator effects. An algorithm detects a moderator effect correctly when the *R2* is greater than zero. In this study, power is defined as the proportion of datasets in which an algorithm had a *R2* > 0.

Variable selection is defined in terms of the algorithms ability to accredit positive variable importance values to relevant moderators. Variable importance measures capture the relative contribution of various moderators. In this study the variable importance measures are extracted, and rescaled to the sum to 100 within each simulated dataset.

Because the used dataset was very large (777600 observations), all p-values were significant. Therefore, we instead focused on the effect size and used partial eta squared .

**Design**

In this study we have manipulated five design factors: the number of studies *k* (20, 40, 80 and 120), the average within-study sample size (40, 80 and 160), the amount of moderators *M* (2-9), the residual heterogeneity *τ2* (0, .01, .04 and .28) and the effect size of the population (.2, .4, .6 and .8). Data were simulated using the random-effects model, based on six models:

(A) Main effect of one moderator,

(B) Two-way interaction,

(C) Three-way interaction,

(D) Two two-way interactions,

(E) Non-linear, cubic relationship,

(F) Exponential relationship,

**Simulation**

This simulation included three versions of MetaForest: 1) random-effects weighted MetaForest, 2) fixed-effects weighted MetaForest, and 3) unweighted MetaForest. Also included were random-effects meta-analysis and the latest published version of metaCART (Li et al., 2019).

A total of 100 datasets were simulated, for every possible combination of the design factors. Each dataset consisted of a training sample, with a number of studies equal to the design factor *k*, and a testing sample of 100 studies. For every study , the values of *M* moderators were random drawn from a skewed normal distribution. The dependent variable represents Hedges’ , an estimator of the standardized mean difference between a treatment condition and a control condition (Harrer et al, 2019). To obtain the true effect size , we started sampling it from a normal distribution for every study. This normal distribution contained a mean that was computed by the evaluation of the model for the value of the effect size , the vector of the predictors and the between-studies variance . Varying the within-study sample size brought in sampling error: every study sample size was drawn from a normal distribution with mean , and standard deviation /3 (Viechtbauer, 2007). The observed effect size was then drawn from a non-central t-distribution, assuming an equal number of cases in the treatment and control group. (Li et al., 2017).

**Results**

All the results can be found in a more extensive and detailed way in: Table 1 and 2, Figures 1 up to and including 6, and Supplementary Figures 1 and 2.

**Absolute performance:** The mean was highest for random-effects meta-analysis (*M* = .55 and *SD* = .28). Nearly similar were random-effects MetaForest (*M* = .45 and *SD* = .22) and unweighted MetaForest (*M* = .45 and *SD* = .19). Followed up by fixed-effects MetaForest (*M* = .39 and *SD* = .19) and at last metaCART (*M* = .28 and *SD* = .31). The partial s of the effects of all design factors on the prediction accuracy are displayed in Table 1.

For the effect-size there were found large effects on every algorithm. For all models, the predictive accuracy increased when the effect size increased (see Figure 1). There were a few exceptions, the predictive accuracy in model C of metaCART and fixed-effects MetaForest decreased when the effects size went from .50 to .80, this was the same case in model D. In model E fixed-effects MetaForest decreased heavily as for metaCART had a slight decrease. In model F only fixed-effects MetaForest decreased after the effect size went from .50 to .80.

The presence of residual variance caused a decrease in for all models. In other words: Decreased when increased (see Figure 2). Random-effects meta-analysis had the highest in all models. Followed by random-effects MetaForest and unweighted MetaForest, which were constantly on a similar level. MetaCART had the lowest in model A up to and including D. The lowest accounted for fixed-effects MetaForest in model E. In model F, both metaCART and fixed-effects MetaForest had the smallest.

The amount of studies *k* also had an effect on the performance of all algorithms. In Figure 3, *k* interacted with . The effects for these combinations of factors were most present for random-effects meta-analysis, and were followed up by random-effects MetaForest and unweighted MetaForest. Next came fixed-effects MetaForest, and at last to a much smaller extent metaCART. This indicates that especially metaCART profits less from more studies *k* when gets larger. Interactions between *k* and the models (see figure 4) showed that in model C and D metaCART needs a certain amount of studies to start explaining variance. For every model, random-effects meta-analysis had the highest . When the amount of studies was low (e.g. 20), metaCART had the lowest. In the linear models A up to and including D Random-Effect Meta-Analysis had the highest for every amount of *k*, this was different for models E and F. In both models, when *k* reached an amount of 120, random-effects MetaForest, unweighted MetaForest, and metaCART ended up eventually higher than random-effects Meta-Analysis. Only fixed-effects MetaForest did not reach a higher .

Additionally, the average within-study sample size was associated with . A larger caused a better predictive accuracy (see Figure 5). This accounted for all algorithms. The effect of the amount of moderators *M* led to the outcome that only random-effects meta-analysis kept relatively even prediction accuracy when *M* got higher (see Figure 6). Random-effects MetaForest and unweighted MetaForest decreased slightly (, as there was the same amount of difference for fixed-effects MetaForest. MetaCART’s prediction accuracy decreased heavily when the amount of moderators went up until 5, but also up surged when the amount of moderators got to 6. After that it kept decreasing when the amount of moderators increased. This indicates that metaCART is most vulnerable, and random-effects meta-analysis least vulnerable, to the presence of irrelevant moderators.

**Relative performance:** The mean performance difference between random-effects meta-analysis and random-effect MetaForest was *M* = .10 and *SD* = .05. Between random-effects MetaForest and metaCART was *M* = .17 and *SD* = .09 and random-effects MetaForest and fixed-effects MetaForest differed slightly with *M* = .06 and *SD* = .03. A negligible difference was present between random-effects MetaForest and unweighted MetaForest (*M* = .00 *SD* = .01).

All of these numbers indicate a superior performance of random-effects meta-analysis and an inferior performance of metaCART over all models. Furthermore, random-effects MetaForest had a considerable advantage over fixed-effects MetaForest and in particular over metaCART. The difference between random-effects MetaForest and unweighted MetaForest is virtually non-present.

For a very slight difference between random-effects MetaForest and unweighted MetaForest, the most important predictors were the effect-size , the model and the residual heterogeneity , and the interaction between those two design factors. In model E and F in Figure 2, unweighted MetaForest outperforms random-effects MetaForest marginally when the effect size got higher. Also in Model E and F in Figure 2, unweighted MetaForest exceeds random-effects MetaForest minimalistic when is present.

For the difference between random-effects meta-analysis, random-effects MetaForest and metaCART, the number of studies *k* and the model, and the interaction between those two factors had a clear considerable effect. Random-effects meta-analysis and random-effects MetaForest perform solid when *k* is low, metaCART on the other hand, needed a certain amount of studies to perform, especially in models A up to and including D. In the non-linear models E and F, metaCART still needed a certain amount of studies to perform, but in these models this effect was less present.

**Power:** The proportion of cases in which each model achieved a positive determined the power of all six models. For the statistical power of these models, we investigated the conditions under which random-effects meta-analysis, random-effects MetaForest and metaCART achieved a positive in at least 80% of the datasets. The full reports are shown in Supplementary Figures 1 and 2.

Supplementary Figure 2 shows the influence of the six design factors that are discussed in the introductory chapter of this result section. These results display the conditions under which random-effects MetaForest and metaCART reach > 80% power. In most conditions, random-effects MetaForest had sufficient power, even when was high and *k* was small. There were some exceptions. For model b, when was low (e.g. 0.2) and was high (e.g. 0.28), it did not matter how many moderators were present. It did not reach > 80% power in any condition. This also accounted for metaCART. In an identical condition, except for a lower (e.g. 0.04), Random-effects MetaForest could only reached > 80% power when the average study sample size was high (e.g. 80 - 160). Another interesting aspect is that metaCART did not reach > 80% power in conditions where random-effects MetaForest did not reach > 80% power. In other words: metaCART only reached > 80% power when random-effects MetaForest also did. However, MetaForest did reach > 80% power numerous times when metaCART did not. This suggests a superior performance of random-effects MetaForest over metaCART.

Supplementary Figure 1 displays the same conditions showed in Supplementary Figure 2, only the comparison is now between random-effects MetaForest and random-effects meta-analysis. Nearly all conditions had sufficient power by both algorithms, but there were some exceptions. Again, for model b, when was low (e.g. 0.2) and was high (e.g. 0.28), it did not matter how many moderators were present. It did not reach > 80% power in any condition. Only random-effects meta-analysis managed to obtain enough power when *k* was high (e.g. 80 - 120) and was high (e.g. 80 – 160). Furthermore, random-effects MetaForest showed a little advantage over random-effects meta-analysis in terms of reaching > 80% power when the amount of studies *k* was low (e.g. 20).

**Variable selection:** To investigate which design factors predicted the standardized variable importance of a relevant moderator we conducted ANOVA’s. All partial s are displayed in Table 2, which considers the design factors and their interactions. The most important design factors for random-effects meta-analysis were the effect size ( and the model (. For random-effects MetaForest these were: effect size (, amount of studies *k* (, the model (, residual heterogeneity ( and the interaction between effect size and model (. At last, for metaCART, the most important design factors were: the model (, amount of studies *k* ( and the effect size .

**Discussion**

This paper evaluated the performance of five algorithms for exploring heterogeneity in meta-analysis. With a simulation study we investigated the predictive performance of random-effects, fixed-effects, and unweighted MetaForest, which all three utilize random-forests algorithm. We also included linear random-effects meta-analysis, and the single tree based algorithm metaCART (Li et al., 2017). In most conditions, random-effects meta-analysis outperformed random-effects, fixed-effects and unweighted MetaForest as well as metaCART. These findings suggest that random-effects meta-analysis is the most suitable technique when it comes to exploring heterogeneity in meta-analysis when moderator effects are non-normally distributed.

Addressing the simulation results more specifically, random-effects meta-analysis indicated the overall best predictive performance, followed by random-effects MetaForest and unweighted MetaForest, whose difference was negligible. Fixed-effects MetaForest showed the third best predictive performance and at last came metaCART. The difference between random-effects MetaForest and random-effects meta-analysis was largely explained by nearly all combinations of design factors, as can be seen in the figures 1 up to and including 6. The most clear performance differences were when: the effect size became bigger (e.g. 0.80); *k* was low (e.g. 20) and when the residual heterogeneity was absent. There were only two occasions in which random-effects meta-analysis did not explained the most variance. This occurred when the marginal interacted with a high number of studies (e.g. *k* = 80 or 120) in model E and F (see Figure 4). In these situations, all three MetaForest algorithms explained more variance than random-effects meta-analysis did. It is important to address that this occurred in the two models that simulated non-linearity.

This study also implied that random-effects meta-analysis and random-effects MetaForest both can have enough power for the identification of relevant moderators. Generally speaking, in most conditions both algorithms reached sufficient power. Even a small amount of 20 studies was in most cases enough. There were some exceptions under a couple of conditions. When; effect size was low, number of studies *k* was low, the average within-study sample size was low, and there was a high amount of moderators M, there were some repeating measures were only MetaForest reached sufficient power. Also in model B, it was difficult for both algorithms to reach sufficient power, especially when the residual heterogeneity was high. It is recommendable to take a look at Supplementary Figure 1 for the full report of the conditions under which random-effects MetaForest and metaCART reached sufficient power. Furthermore it became clear that linear models and random forests have an advantage over single trees in terms of reaching sufficient power when study effects are non-normally distributed. In Supplementary Figure 2 and 3 it is evident that the identification of relevant moderators capabilities for metaCART where much smaller than for random-effects meta-analysis and random-effects MetaForest. MetaCART only reached sufficient power when random-effects MetaForest and random-effects meta-analysis also did, it did not had a single circumstance in which it was the only algorithm that reached sufficient power.

The most noteworthy aspect after interpreting the results of this study is that random-effects meta-analysis, a linear relationship, had an overall better performance than random-effects MetaForest, a random forests algorithm. We did not expected this in our hypotheses, as we expected random-effects MetaForest to perform advantageously over all other algorithms. Namely because of: the ability of random-forests to handle interactions between moderators and non-linear effects, their greater flexibility in modeling complex datasets than linear models do (Earp & Trafimow, 2015). And the real-data examples of simulations studies that indicated that even metaCART was suggested as an useful tool for detecting interactions between moderators in meta-analysis (Dusseldorp et al., 2014; Li, Dusseldorp & Meulman, 2017). A notable remark hereby is the exception when MetaForest actually did explained more variance than random-effects meta-analysis in models E and F (see Figure 4). These were the non-linear distributed simulated datasets, in which MetaForest outperformed random-effects meta-analysis when the amount of studies was high (e.g. 80 or 120).

Obviously, the main reason why random-effects meta-analysis performed better than all other algorithms was because the model was generated by the same modeling technique as the reference model. Van de Ploeg, Austin and Steyenberg (2014) call this a ‘home advantage’ over models generated by a different modeling technique than the reference model. After they became aware about this ‘home advantage’, they provided the performance of models according to different reference models for a fair estimate of the performance of the considered approaches. Despite the fact that we did not do this in our study, we will take a look when linear models could perform better than random-forests.

Linear models will perform better when the underlying data generating function is truly linear (Smith, Ganesh & Liu, 2013). Estimating such processes with a linear function will fit the data very efficiently. Random forests, in this case, would approximate the linear function, but not exactly (Grömping, 2009). In this study we indeed saw this occur in model A up to and including D (see Figure 4). In the non-linear shaped models E and F, it became clear that for linear models it was more difficult to capture the non-linear features when the amount of studies was high (e.g. 80 or 120). Eventually all MetaForest algorithms explained more variance in these conditions. This also may support the theory that linear models need less data to get good results than random-forests (Caruana & Niculescu-Mizil, 2006). Additionally, van der Ploeg, Austin and Steyerberg (2014) concluded that linear models are especially useful in relatively small data sets. With very small data sets there is not any modeling technique that will perform satisfactory, but their results confirmed that because of the increased flexibility of random-forests larger sample sizes are required for a reliable estimation. Which can explain why in this study random-forests performed better when the dataset became considerably large.

The cause of this could be because random-forests algorithms do not use the data as efficiently as conventional techniques like linear models do, in other words, they are more ‘data hungry’. Van der Ploeg, Austin and Steyerberg (2014) compared techniques like random-forests, singletree based algorithms (CART) and linear models. They figured that random-forests needed far more events per variable to achieve a stable optimism (difference between training error and in-sample error) of < 0.01 and an approved area under the curve (AUC) than linear models and CART. AUC is used in classification analysis in order to determine which of the used models predicts the classes as best; models with higher AUC’s are favored over models with lower AUC’s. Widely used AUC applications are ROC-curves (Gravetter & Wallnau, 2016). The CART model had a stable performance, but this was in comparison with the other techniques at a relatively low level, like metaCART in this study. Van der Ploeg, Austin and Steyerberg (2014) implied when very large data sets are available; random-forests may achieve an AUC-value that is superior of modeling techniques like linear models. Which indicates that random-forests should only be considered if very large data sets with many events are available.

Another reason why a linear model would perform better than random-forests algorithms is because no extrapolation is possible (Hengl et al., 2018). MetaForest and metaCART, and every other decision-tree based approach, identify homogeneous areas in the data and assign the mean value of the data that is incorporated in those areas to the corresponding ‘leaf’. Therefore, they are so-called ‘granular’ and then, they must show a series of steps in the outputs. The algorithms that are based on random-forests do not show that phenomenon clearly, it gets nuanced because of the large number of trees. If a certain value is exterior of the original range, that datum is assigned to the ‘leaf’ that included the extreme condition that was found in the training dataset. This results in an output that is the mean value of the values contained in that leave (Hengl et al., 2018). Thus, extrapolation is not possible.

Altogether, the random-effects meta-analysis model could be very useful or even the most appropriate technique to choose in a couple of situations. When the underlying data generating function of the dataset is linear, it will fit the data more efficiently than the approximation of tree-based methods. There is also an indication that linear models should perform better when the dataset is relatively small, and tree-based algorithms are favorable in larger sized datasets. This can be caused due to the data ‘hungriness’ of the latter techniques. Also, when there is a need to extrapolate, random-forests is not an option, in this case only linear models will work.

To compare the tree-based methods, all three MetaForest algorithms had a superior performance over metaCART. Especially random-effects MetaForest and unweighted MetaForest, who were on the same level on nearly all occasions. Also fixed-effects MetaForest performed better than metaCART. After concluding that MetaForest was a superior tool over metaCART for exploring heterogeneity when study effects were normally distributed (van Lissa, 2017). Now these findings suggest that when it comes down to non-normally distributed data; MetaForest is the in favorable technique as well.

This is because random forests overcome many in limitations of single trees (Breiman, 2001). Since they are already mentioned in the introduction, we will sum them up. As a result of all grown individual trees on bootstrapped samples of the dataset it will unveil variance between all trees. Only small random selections of moderators are contemplated as splitting variables, thus this variance will increase. The predictions of all those trees will then be averaged. By doing this, random-forests turn the instability of individual trees into an advantage (van Lissa, 2017). In the end, this allows for the detection of correlated predictors and smaller effects that would otherwise wear out in a single tree. Also singletree-based approaches have a higher tendency to overfit, because there are multiple trees in random-forests the tendencies to overfitting decreases due to bagging and random feature selection. Hence, after a certain number of trees the performance tends to stay in a certain value (Hastie, Tibshirani & Friedman, 2009). Further, the predictions of every tree tend to be a little different. By capturing continuous curvi- linear relationships as well as (complex) interactions much better, random-forests produces better piecewise linear predictions than single trees based approaches are able to (Bühlmann & Yu, 2002). This results in a higher from random-forests and a lower from single tree approaches. We also saw this in the results of this study. At last, in random-forests, the out-of-bag (OOB) error closely approximates the cross-validation error, and provides good estimates of the prediction accuracy in prospective samples (Saffari et al., 2009).

**Strengths, limitations and recommendations**

To begin with a minor limitation of this paper, we did not take the ‘home advantage’ effect into account (van der Ploeg, Austin & Steyenberg, 2014). Generating models by the same modeling technique as the reference model generally perform better over models that are generated with a different technique than the reference model. It is a common in nowadays science to make such an assumption: the training and test data that are represented by the same features and drawn from the same distribution. Also, the performance of these algorithms relies on collecting high quality and sufficient labeled training data to train a statistical or computation model to make prediction on the future data. But, in most real-world scenarios, labeled training data are limited in supply or can only be obtained with high costs. This is a problem that makes it hard to see if machine-learning methods are really applicable in practice (Aggarwal, 2015). Thus, to make an even better comparison between random-effects meta-analysis, MetaForest and metaCART, we should not only provide the performance of the models according to different reference models, for an unbiased assessment in future research. But also, as an addition to the simulation data, a study with using actual published meta-analytical data. It is interesting and important to see how primary input data gets classified by the different algorithms, and how it will contrast with the original classification.

To be able to have a better example when linear models perform in comparison with tree-based approaches, a simulation or real study (ideally both) with a relatively small example should be used. As we discussed earlier, there are numerous examples why linear models could perform better than tree-based approaches if the dataset is small. It is possible that a large sample size is a requirement of tree-based approaches to produce a reliable estimation, which can be due to the high data hungriness standards (van der Ploeg, Austin & Steyerberg, 2014). It would be really relevant for scientists to know when they should use a linear model or a tree-based approach when they are exploring heterogeneity in meta-analysis. Therefore, it would be of great interest to investigate this aspect in future research.

A strength of this study is that we examined the performance of the algorithms in the presence of skewed distributed moderators. That said we give an addition to the research of van Lissa (2017) by making the simulations more representative. Because it is entirely possible in the social sciences that covariate features affect the procedures.

At last, we were able to give an unbiased demonstration that random-effects MetaForest and unweighted MetaForest had a superior performance over fixed-effects MetaForest and MetaCART. In addition to the, in random-effects MetaForest favourable, outcomes of van Lissa (2017), it shows that for now, random-effects MetaForest is also the best tool for exploring heterogeneity in meta-analysis when moderators are skewed distributed.

**Conclusion**

This study has indicated that MetaForest is a potent algorithm for exploring heterogeneity in meta-analysis. Within a large set of possible competitors, MetaForest is able to identify the relevant non-normally distributed moderators. This also accounts for when the number of studies is low, MetaForest frequently showed sufficient power when there were only 20 studies. In addition to the, in random-effects MetaForest favourable, outcomes of van Lissa (2017), it shows that random-effects MetaForest is also the best tool for exploring heterogeneity when moderators are skewed distributed. In further research, it is important to be aware of the ‘home advantage’ effect. For an unbiased assessment, meta-analytical simulation studies should provide the performance of the models according to different reference models. Also, to make an even more representative conclusion, future research should evaluate the algorithms performances with relatively small datasets and actual published meta-analytical data.

Table 1

*Partial eta squared for the influence of design factors on* .

|  |
| --- |
|  |

MF-random MF-FE Uniform MetaCART RMA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.20 | 0.10 | 0.20 | 0.07 | 0.22 |
|  | 0.07 | 0.07 | 0.09 | 0.05 | 0.08 |
|  | 0.17 | 0.19 | 0.17 | 0.16 | 0.03 |
|  | 0.02 | 0.02 | 0.02 | 0.01 | 0.02 |
|  | 0.03 | 0.04 | 0.03 | 0.16 | 0.01 |
|  | 0.10 | 0.04 | 0.10 | 0.26 | 0.05 |
|  | 0.01 | 0.02 | 0.01 | 0.00 | 0.02 |
|  | 0.01 | 0.01 | 0.01 | 0.00 | 0.02 |
|  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
|  | 0.01 | 0.02 | 0.01 | 0.02 | 0.01 |
|  | 0.08 | 0.16 | 0.08 | 0.07 | 0.08 |
|  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
|  | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
|  | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 |
|  | 0.02 | 0.02 | 0.02 | 0.02 | 0.02 |
| *k:n* | 0.01 | 0.01 | 0.00 | 0.00 | 0.01 |
| *k:M* | 0.00 | 0.00 | 0.00 | 0.01 | 0.01 |
| *k:Model* | 0.00 | 0.00 | 0.00 | 0.01 | 0.00 |
| *n:M* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *n:Model* | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 |
| *M:Model* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |

Table 2

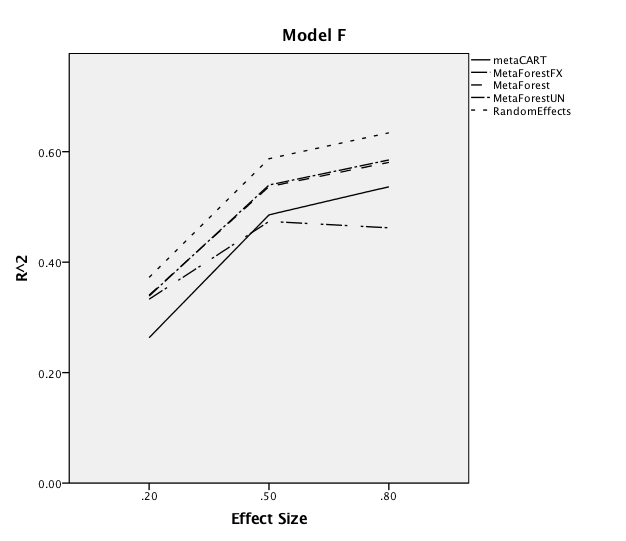
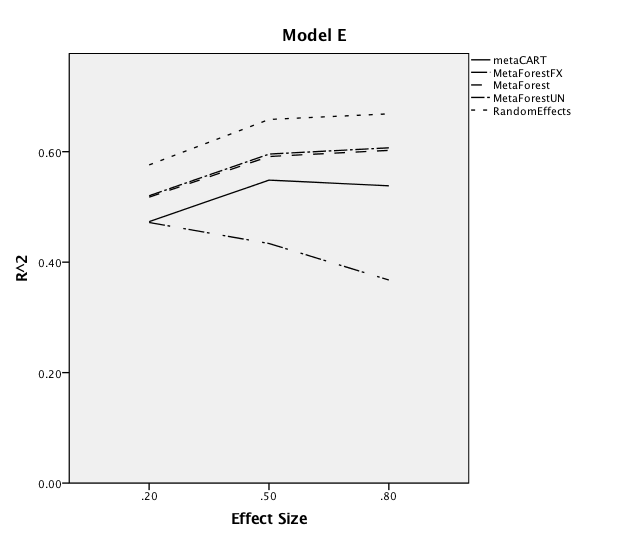
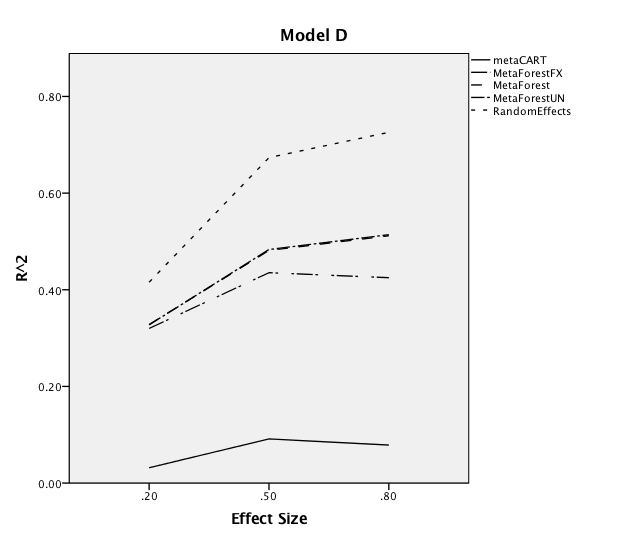
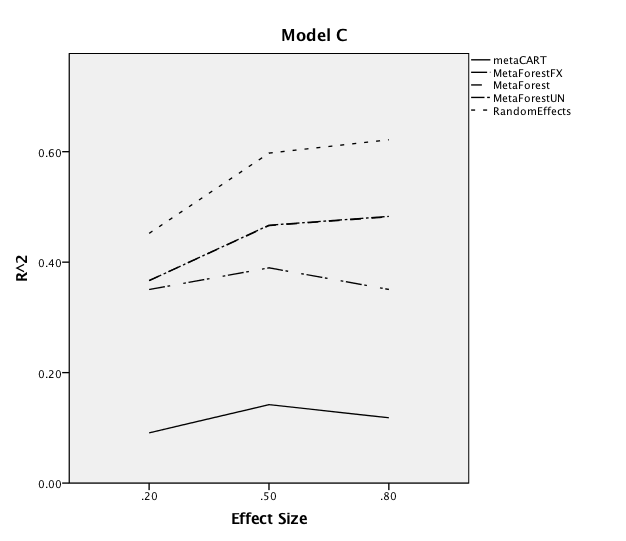
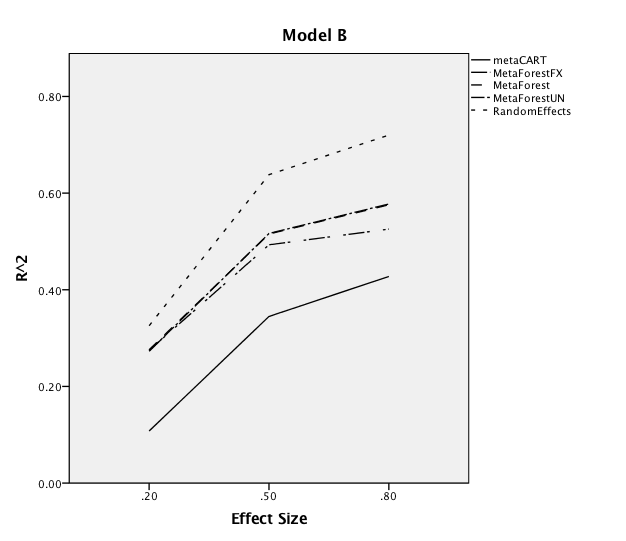
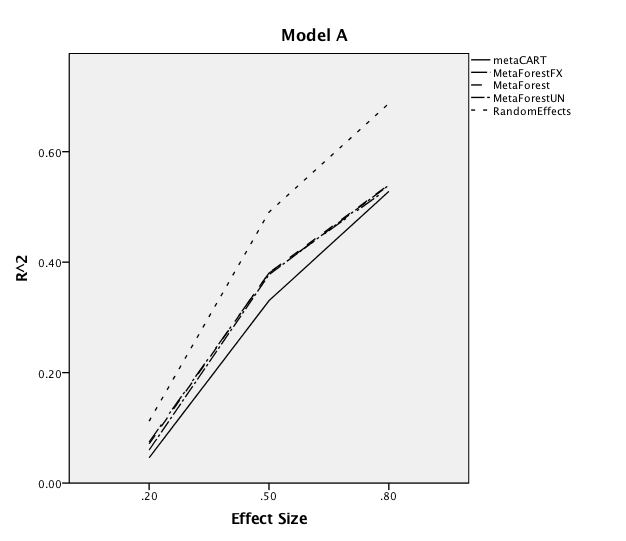
*Partial eta squared for the most important design factors and interactions.*

|  |
| --- |
|  |

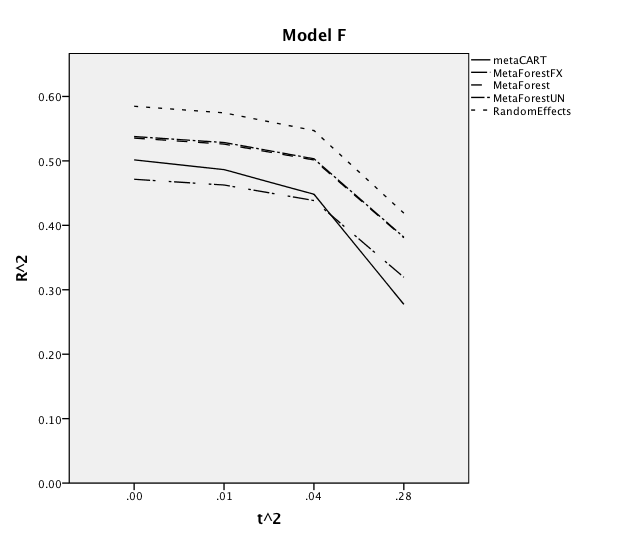
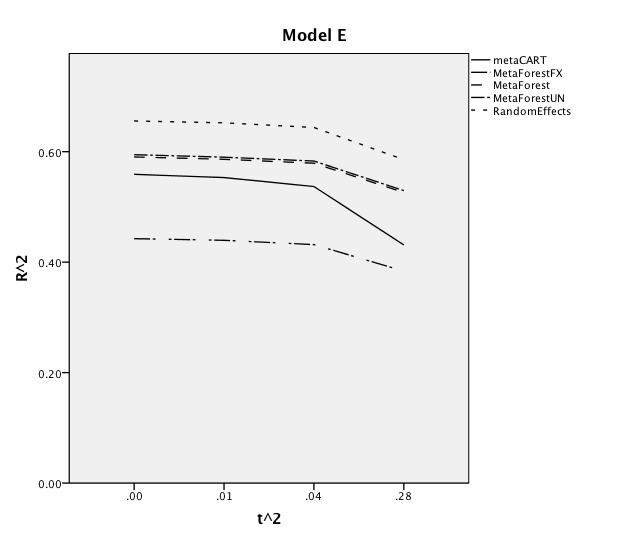
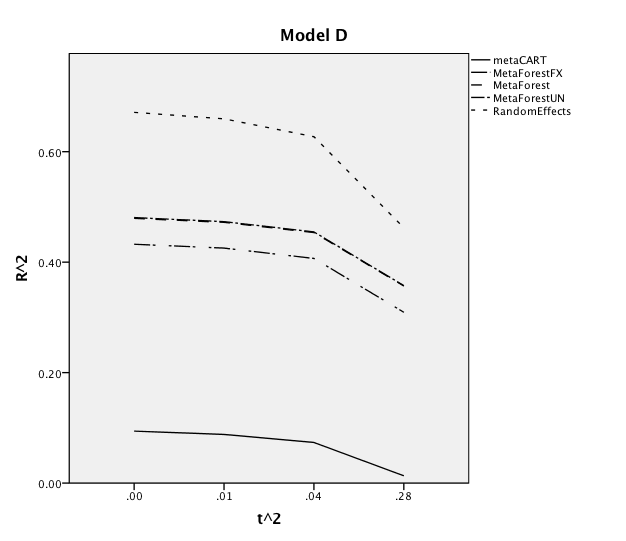
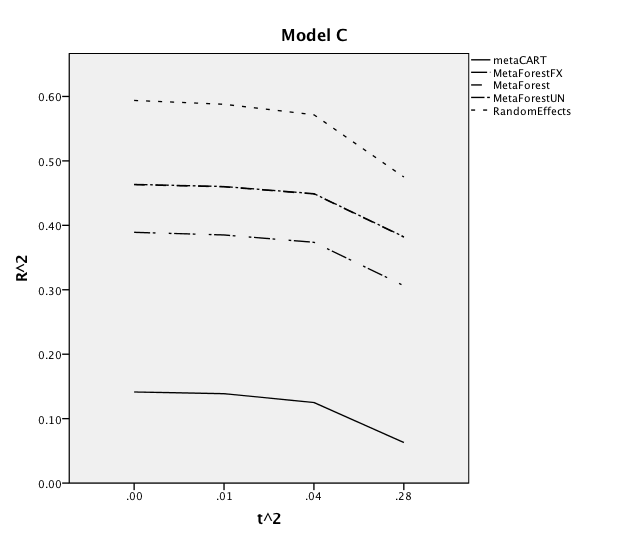
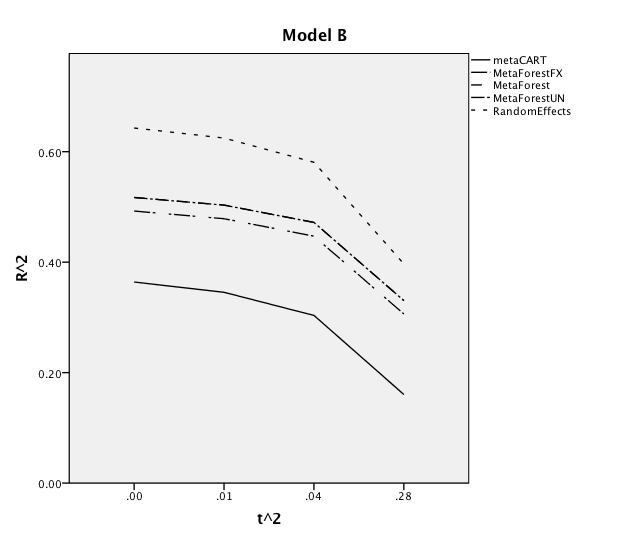
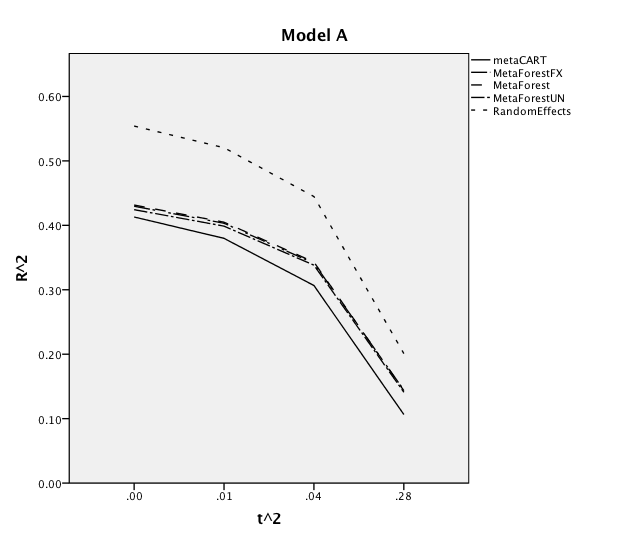
MF-random MF-FE Uniform MetaCART RMA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 0.25 | 0.13 | 0.24 | 0.10 | 0.20 |
|  | 0.11 | 0.10 | 0.10 | 0.06 | 0.08 |
|  | 0.24 | 0.24 | 0.23 | 0.19 | 0.05 |
|  | 0.03 | 0.03 | 0.03 | 0.02 | 0.02 |
|  | 0.09 | 0.11 | 0.09 | 0.01 | 0.02 |
|  | 0.25 | 0.10 | 0.25 | 0.23 | 0.10 |
|  | 0.03 | 0.04 | 0.03 | 0.00 | 0.02 |
|  | 0.02 | 0.01 | 0.01 | 0.03 | 0.01 |
|  | 0.01 | 0.01 | 0.01 | 0.00 | 0.01 |
|  | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |
|  | 0.14 | 0.28 | 0.14 | 0.09 | 0.00 |
|  | 0.01 | 0.01 | 0.01 | 0.01 | 0.00 |
|  | 0.01 | 0.01 | 0.01 | 0.01 | 0.01 |
|  | 0.00 | 0.00 | 0.01 | 0.00 | 0.00 |
|  | 0.05 | 0.05 | 0.04 | 0.02 | 0.03 |
| *k:n* | 0.00 | 0.01 | 0.00 | 0.00 | 0.00 |
| *k:M* | 0.00 | 0.00 | 0.00 | 0.00 | 0.02 |
| *k:Model* | 0.03 | 0.02 | 0.03 | 0.03 | 0.00 |
| *n:M* | 0.00 | 0.00 | 0.00 | 0.00 | 0.00 |
| *n:Model* | 0.01 | 0.02 | 0.01 | 0.01 | 0.01 |
| *M:Model* | 0.01 | 0.01 | 0.01 | 0.00 | 0.00 |

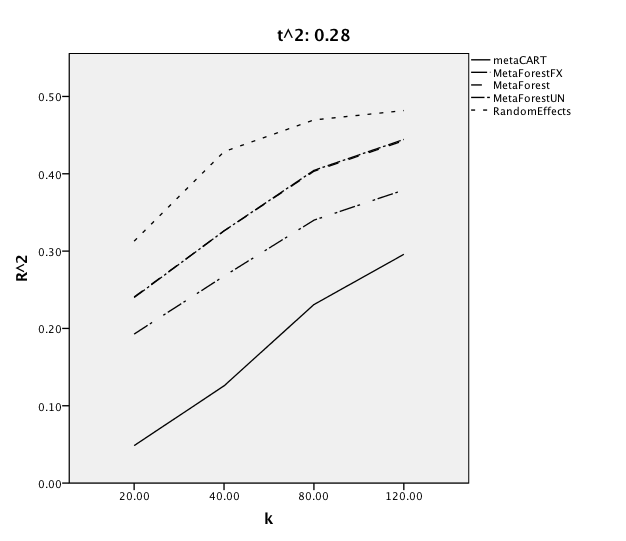
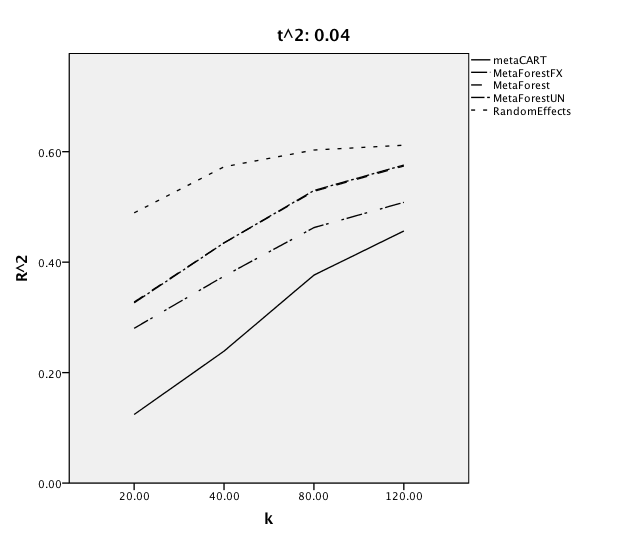
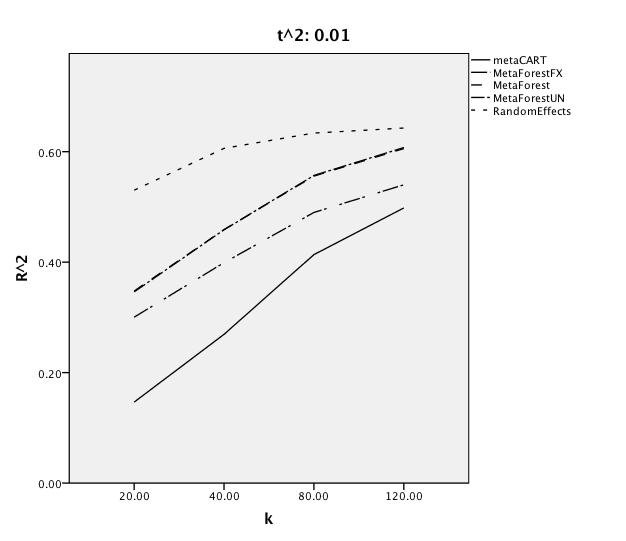
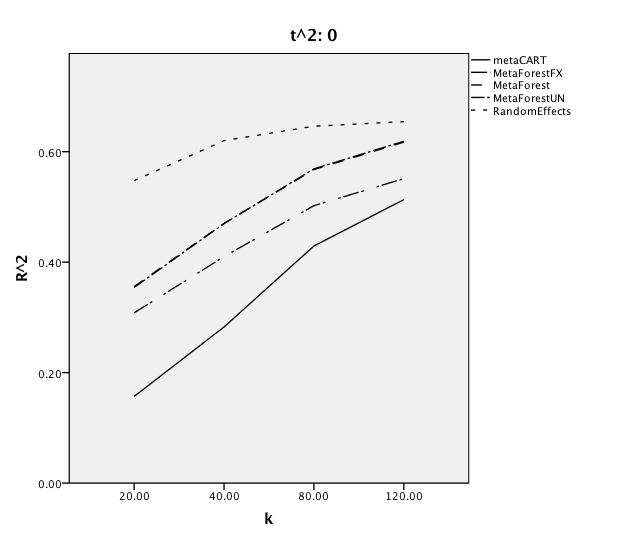
*Figure 1.* Marginal for the interaction between effect size and model



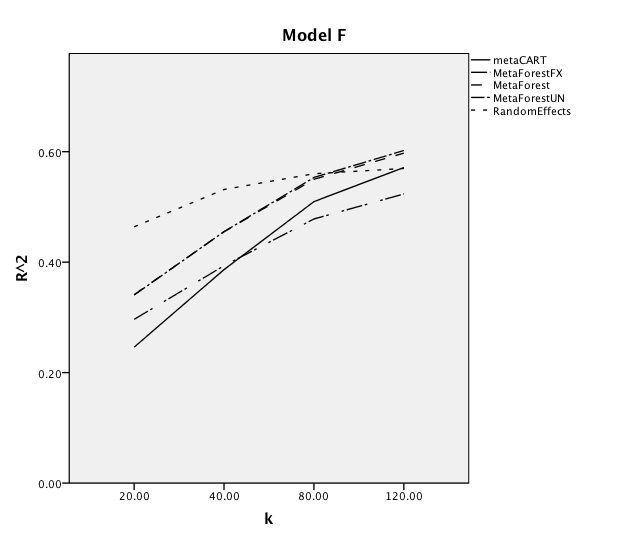
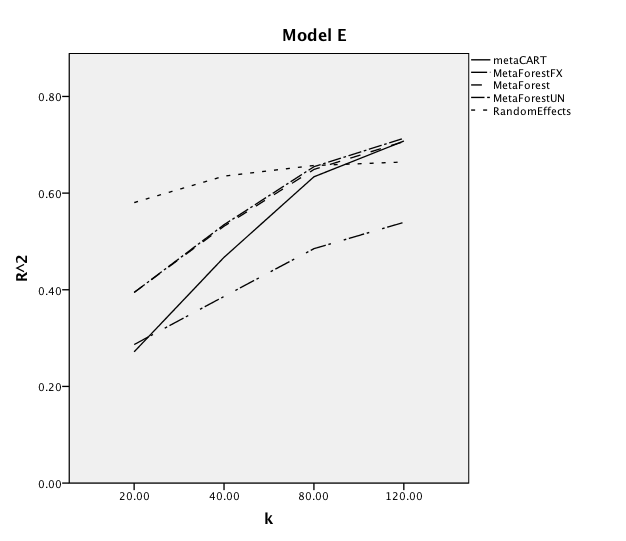
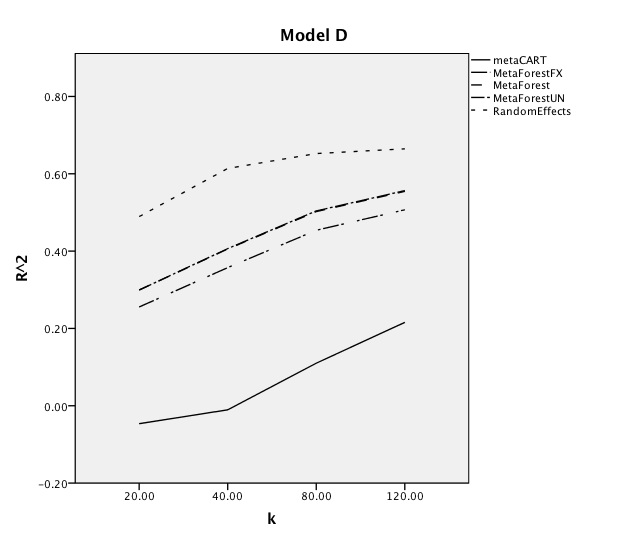
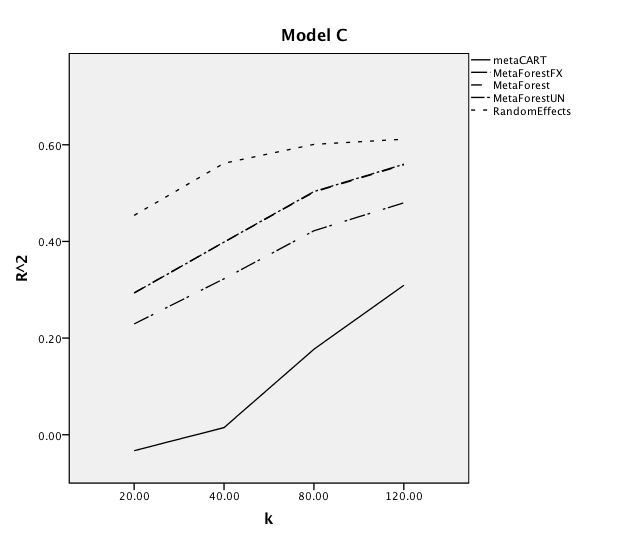
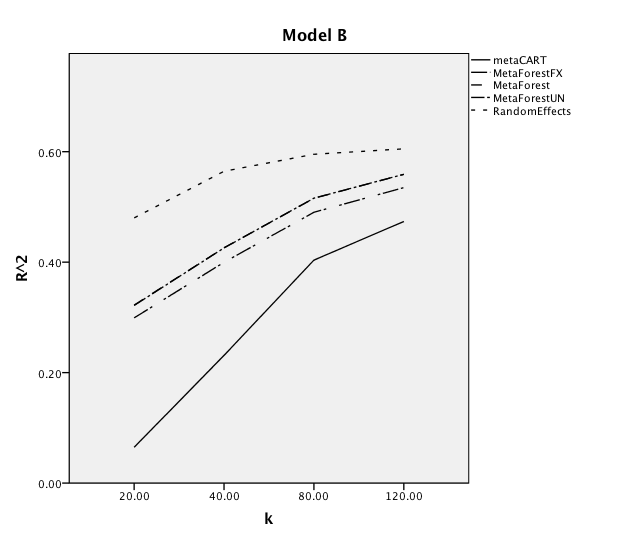
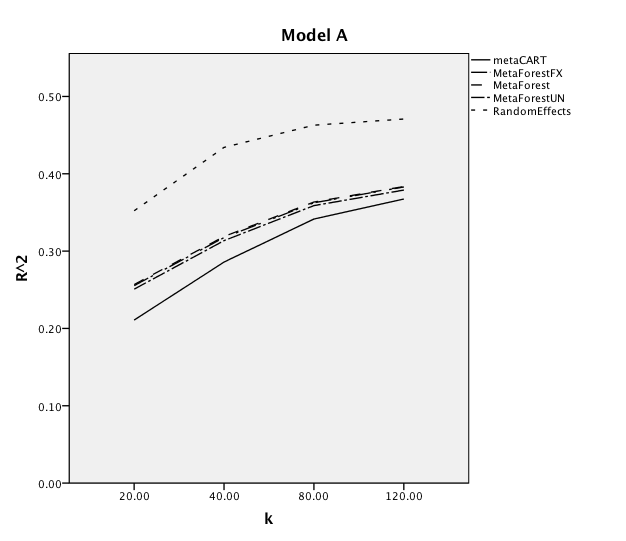
*Figure 2.* Marginal for the interaction between residual heterogeneity and the model



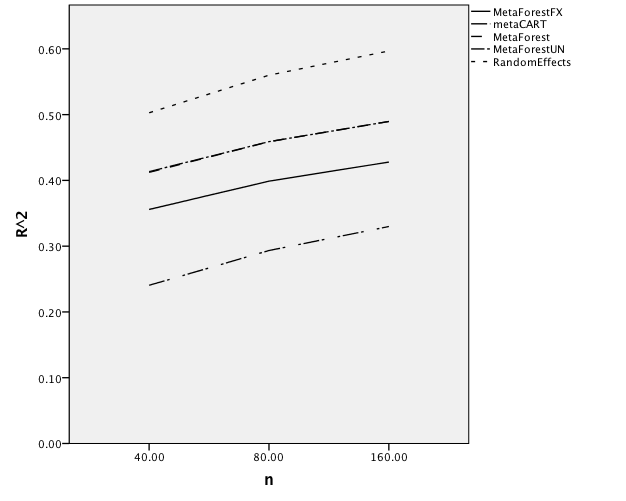
*Figure 3.* Marginal for the interaction between the number of studies and residual heterogeneity



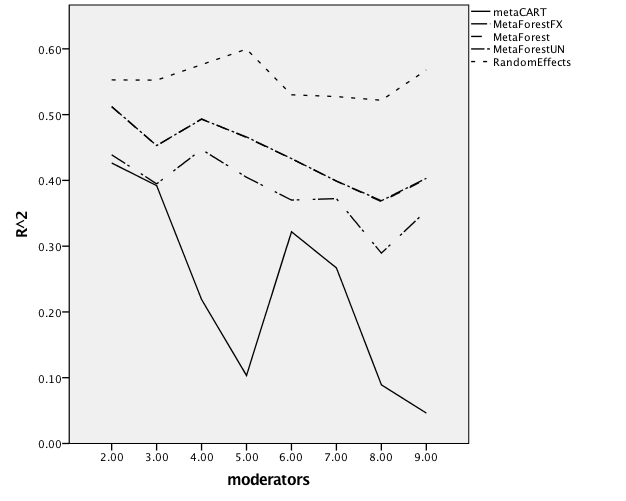
*Figure 4.* Marginal for the interaction between the number of studies and the model



*Figure 5*. Marginal for the effect of the average within-study sample size



*Figure 6.* Marginal for the effect of the amount of moderators



**References**

Aggarwal, C. (2015). *Data Classification: Algorithms and Applications* (2nd edition). consulted from. https://doc.lagout.org/science/0\_Computer%20Science/2\_Algorithms/Data%20Classification\_%20Algorithms%20and%20Applications%20%5BAggarwal%202014-07-25%5D.pdf

Aronson, E., Wilson, T. D., & Akert, R. (2016). *Social Psychology, Global Edition* (9th edition). London, England: Pearson Education Limited.

Auerswald, M., & Moshagen, M. (2015). Generating Correlated, Non-normally Distributed Data Using a Non-linear Structural Model. *Psychometrika*, *80*(4), 920–937. <https://doi.org/10.1007/s11336-015-9468-7>

Bohning, D, Malzahn, U, Dietz, E, Schlattmann, P, Viwatwongkasem, C, Biggeri, A. (2002). Some general points in estimating heterogeneity variance with the DerSimonian-Laird estimator. *Biostatistics*, 3(4): 445–457.

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

Brockwell, SE, Gordon, IR. (2001). A comparison of statistical methods for meta-analysis. *Statistical Medicine*, 20(6): 825–840.

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

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

Caruana, R., & Niculescu-Mizil, A. (2006). An empirical comparison of supervised learning algorithms. *Proceedings of the 23rd international conference on Machine learning - ICML ’06*, *5*(1), 44–54. https://doi.org/10.1145/1143844.1143865

Cesario, J. (2014, January). Priming, Replication, and the Hardest Science. Perspectives

on Psychological Science, 9 (1), 40–48. doi: 10.1177/1745691613513470

Chen, H., Manning, A., & Dupuis, J. (2012). A method of moments estimator for random effect multivariate meta-analysis. *Biometrics*, *68*, 1278–1284. https://doi.org/10.1111/j.1541-0420.2012.01761

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

Dwyer, K., & Holte, R. (2007). Decision Tree Instability and Active Learning. *Machine Learning: ECML 2007*, *34*, 128–139. https://doi.org/10.1007/978-3-540-74958-5\_15

Earp, B. D., & Trafimow, D. (2015, May). Replication, falsification, and the crisis of

confidence in social psychology. *Frontiers in Psychology*, 6 . doi:

10.3389/fpsyg.2015.00621

Follmann, D. A., & Proschan, M. A. (1999). Valid Inference in Random Effects Meta-Analysis. *Biometrics*, *55*(3), 732–737. <https://doi.org/10.1111/j.0006-341x.1999.00732.x>

Genuer, R., Poggi, J.-M., & Tuleau-Malot, C. (2010). Variable selection using random forests. *Pattern Recognition Letters*, *31*(14), 2225–2236. https://doi.org/10.1016/j.patrec.2010.03.014

Gravetter, F., & Wallnau, L. (2016). *Statistics for The Behavioral Sciences* (10th edition). ‎Boston, Massachusetts: Cengage Learning Inc.

Grömping, U. (2009). Variable Importance Assessment in Regression: Linear Regression versus Random Forest. *The American Statistician*, *63*(4), 308–319. https://doi.org/10.1198/tast.2009.08199

Harrer, M., Cuijpers, P., Furukawa, T.A, & Ebert, D. D. (2019). *Doing Meta-Analysis in R: A Hands-on* Guide. DOI: 10.5281/zenodo.2551803.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The elements of statistical learning:

Data mining, inference, and prediction (Second ed.). New York: Springer.

Hedges, L. V., & Vevea, J. L. (1998). Fixed- and Random-effects Models in

Meta-analysis. Psychological Methods, 3 (4), 486–504.

Ioannidis, J. P. A. (2016). Why Most Clinical Research Is Not Useful. *PLOS Medicine*, *13*(6), 33–44. https://doi.org/10.1371/journal.pmed.1002049

Jackson, D., & White, I. R. (2018). When should meta-analysis avoid making hidden normality assumptions? *Biometrical Journal*, *60*(6), 1040–1058. https://doi.org/10.1002/bimj.201800071

Kontopantelis, E., & Reeves, D. (2010). Performance of statistical methods for meta-analysis when true study effects are non-normally distributed: A simulation study. *Statistical Methods in Medical Research*, *21*(4), 409–426. https://doi.org/10.1177/0962280210392008

Kunert, R. (2016). Internal conceptual replications do not increase independent replication success. *Psychonomic Bulletin & Review*, *23*, 1631–1638. https://doi.org/10.3758/s13423-016-1030-9

Laird, N. (1978). Nonparametric Maximum Likelihood Estimation of a Mixing Distribution. *Journal of the American Statistical Association*, *73*(364), 805–811. Geraadpleegd van https://amstat.tandfonline.com/doi/abs/10.1080/01621459.1978.10480103?journalCode=uasa20#.Xfva6r97muU

Li, X., Dusseldorp, E., & Meulman, 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. https://doi.org/10.1111/bmsp.12088

Micceri, T. (1989). The unicorn, the normal curve, and other improbable creatures. *Psychology Bulletin*,105(1): 156–166

Morris, T. P., White, I. R., & Crowther, M. J. (2019). Using simulation studies to evaluate statistical methods. *Statistics in Medicine*, *38*(11), 2074–2102. <https://doi.org/10.1002/sim.8086>

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011, February). Interpretation of

random effects meta-analyses. BMJ, 342 , d549. doi: 10.1136/bmj.d549

Saffari, A., Silver, M. J., Zavattari, P., Moi, L., Columbano, A., Meaburn, E. L., & Dudbridge, F. (2017). Estimation of a significance threshold for epigenome-wide association studies. *Genetic Epidemiology*, *42*(1), 20–33. https://doi.org/10.1002/gepi.22086

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. https://doi.org/10.1348/000711007x255327

Sidik, K., & Jonkman, J. N. (2007). A comparison of heterogeneity variance estimators in combining results of studies. *Statistics in Medicine*, *26*(9), 1964–1981. https://doi.org/10.1002/sim.2688

Smith, P. F., Ganesh, S., & Liu, P. (2013). A comparison of random forest regression and multiple linear regression for prediction in neuroscience. *Journal of Neuroscience Methods*, *220*(1), 85–91. https://doi.org/10.1016/j.jneumeth.2013.08.024

Snijders, T. A. B. (2005). Power and Sample Size in Multilevel Linear Models. *Encyclopedia of Statistics in Behavioral Science*, *65*, 32–39. https://doi.org/10.1002/0470013192.bsa492

Steyerberg, E. W. (2019). *Clinical Prediction Models: A Practical Approach to Development, Validation, and Updating* [Ebook] (2nd edition). Geraadpleegd van https://www.springer.com/gp/book/9783030163983

Tatsioni, A., & Ioannidis, J. P. A. (2008). Meta-Analysis. *International Encyclopedia of Public Health*, *5*, 442–450. https://doi.org/10.1016/b978-012373960-5.00341-5

Thompson, S. G., & Higgins, J. P. T. (2002). How should meta-regression analyses be undertaken and interpreted? *Statistics in Medicine,* 21(11), 1559-1573. Doi: 10.1002/sim.1187

Van der Ploeg, T., Austin, P.C. & Steyerberg, E.W. Modern modelling techniques are data hungry: a simulation study for predicting dichotomous endpoints. *BMC Med Res Methodol* **14,** 137 (2014) doi:10.1186/1471-2288-14-137

Van Lissa, C. (2017). *MetaForest: Exploring heterogeneity in meta-analysis using random forests.* Article in preparation.

Viechtbauer, W. (2006). Confidence intervals for the amount of heterogeneity in meta-analysis. *Statistics in Medicine*, *26*(1), 37–52. https://doi.org/10.1002/sim.2514