**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 regression model (Gravetter & Wallnau, 2016).

(1)

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 exceedingly extensive (777600 observations), all p-values were significant. Therefore, we focused on the effect size and used the partial eta squared instead.

**Design**

In this study we have manipulated six 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:

1. Main effect of one moderator,
2. Two-way interaction,
3. Three-way interaction,
4. Two two-way interactions,
5. Non-linear, cubic relationship,
6. 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 metaCART and Random-Effects Meta-Analysis. The in this study included version of metaCART was also used by Li, Dusseldorp & Meulman (2017). Which uses random-effects weights with a pre-specified value of *τ2*, and controls overfitting by pruning the tree until cross-validation error are within the minimum cross-validation error with a standard error of 0.5.

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 serves as Hedges’ , an estimator of the standardized mean difference between a treatment condition and a control condition (Harrer et al, 2019). The true effect size was first sampled from a normal distribution for every study, whereas the mean got computed by evaluating the model for the value of the effect size , the vector of the predictors and the residual heterogeneity . Varying the within-study sample size included 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 Uniform 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 (see figure 2). Decreased when increased. It applies to all models that the effect was obviously strongest for Random-Effects Meta-Analysis. Followed by Random-Effects MetaForest and Unweighted MetaForest, that could constantly be found on a similar level. In model A up to and including D, metaCART had the lowest effect, in model E Fixed-Effects MetaForest, and in model F it accounted for both metaCART and Fixed-Effects MetaForest.

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 exten 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 a 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-Effect 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 has 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-Effect 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 .

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

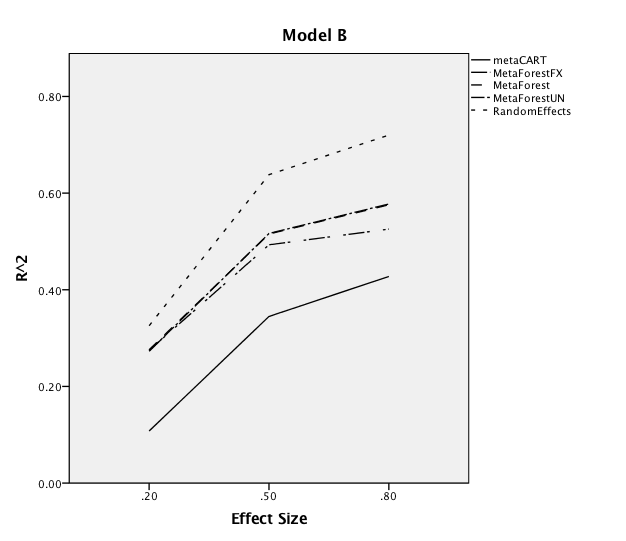
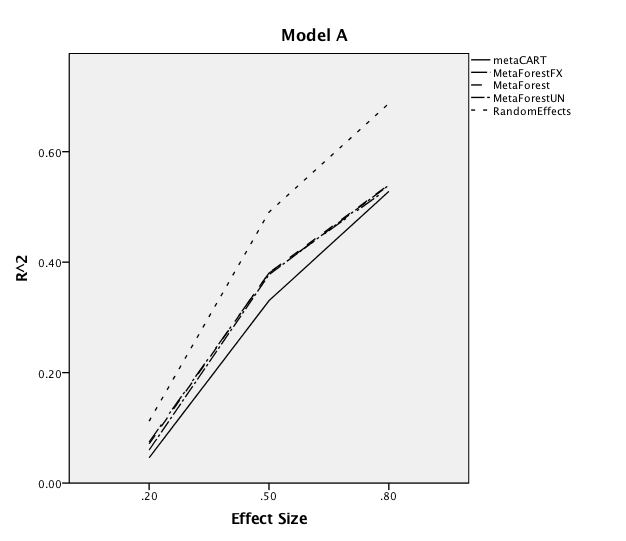
*Partial eta squared for the most important design factors and interactions.*

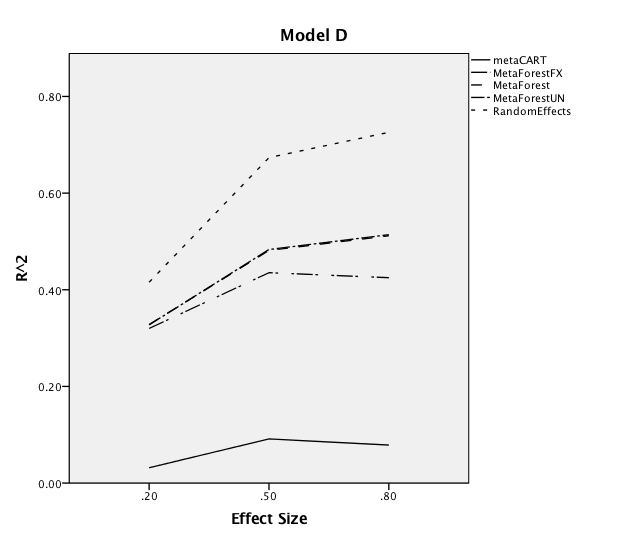
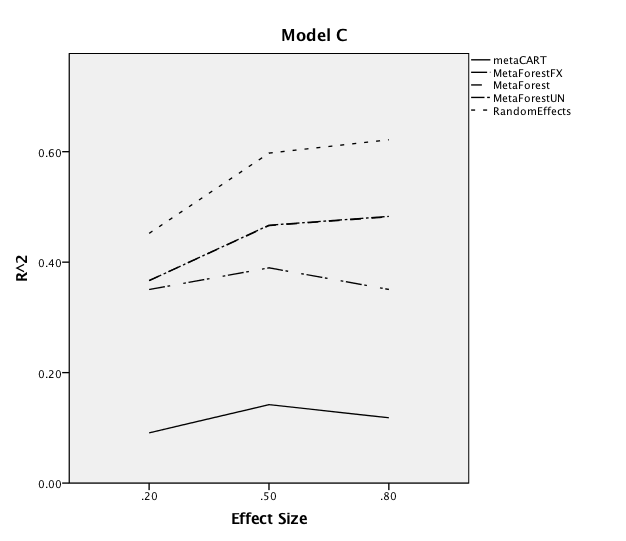
|  |
| --- |
|  |

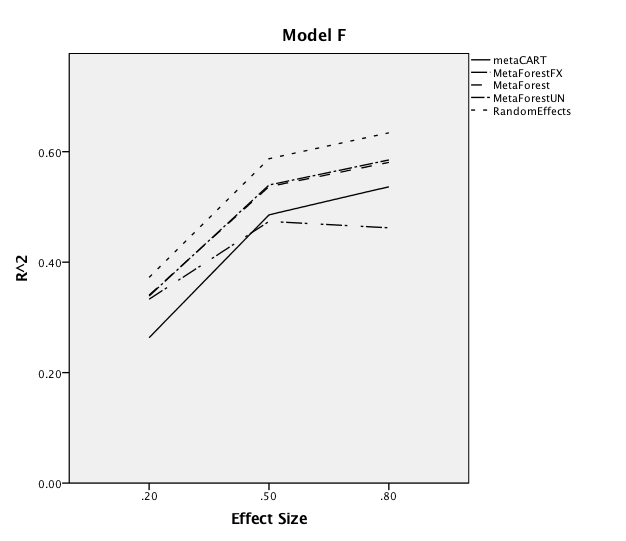
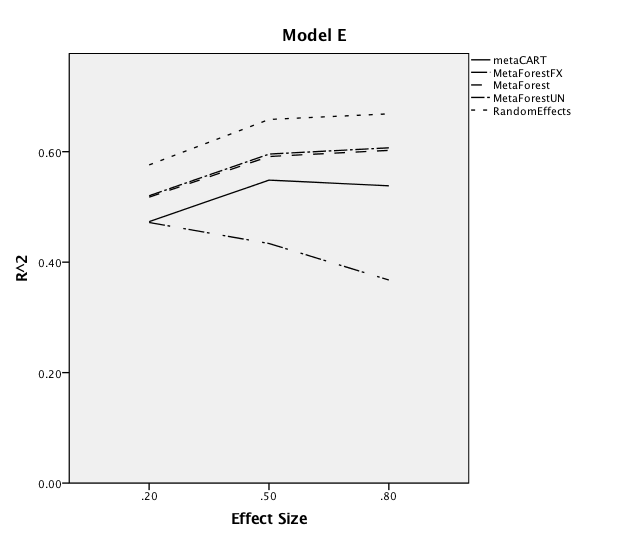
MF-random MF-FE Uniform MetaCART RMA

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | 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 |

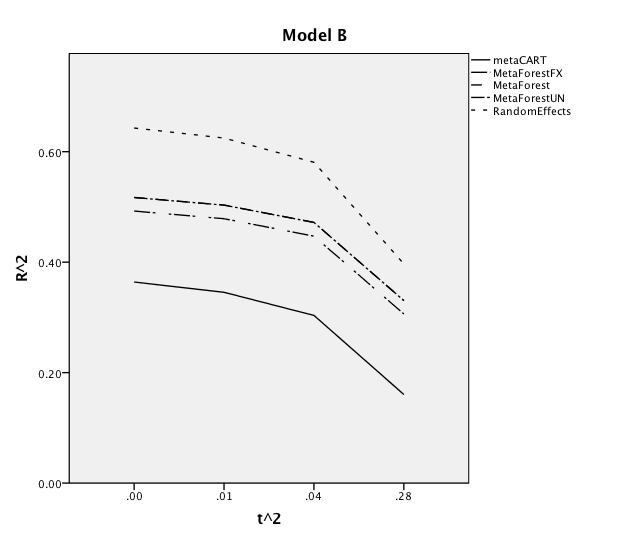
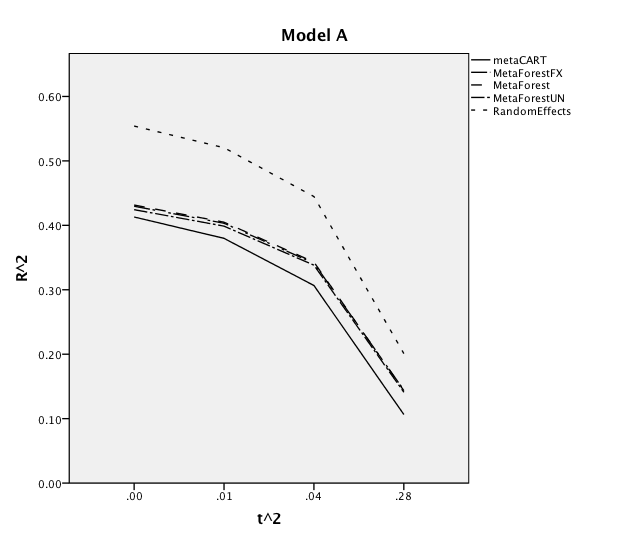
*Figure 1.* Marginal for the interaction between effect size and model

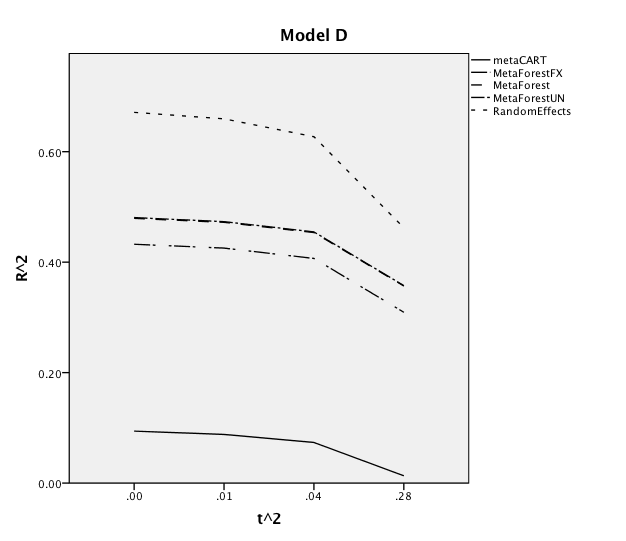
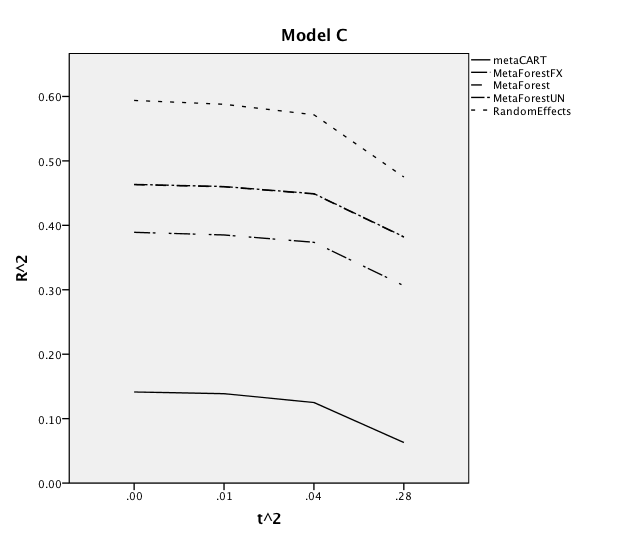


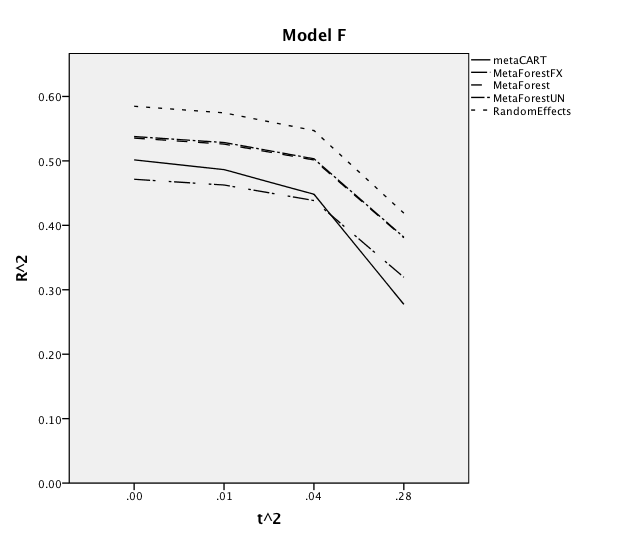
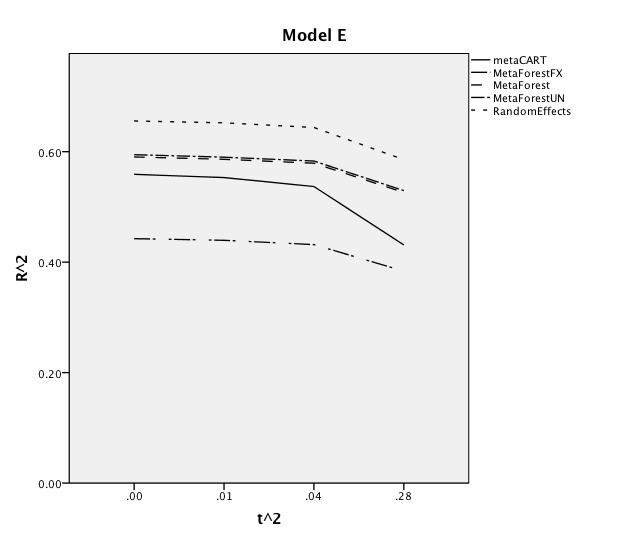




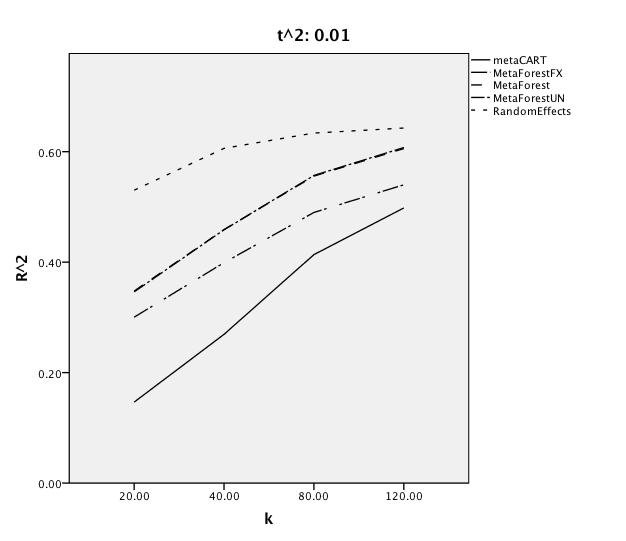
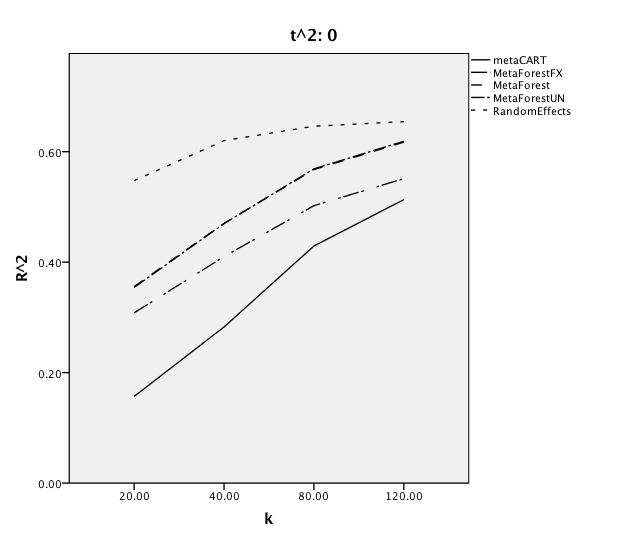
*Figure 2.* Marginal for the interaction between residual heterogeneity and the model

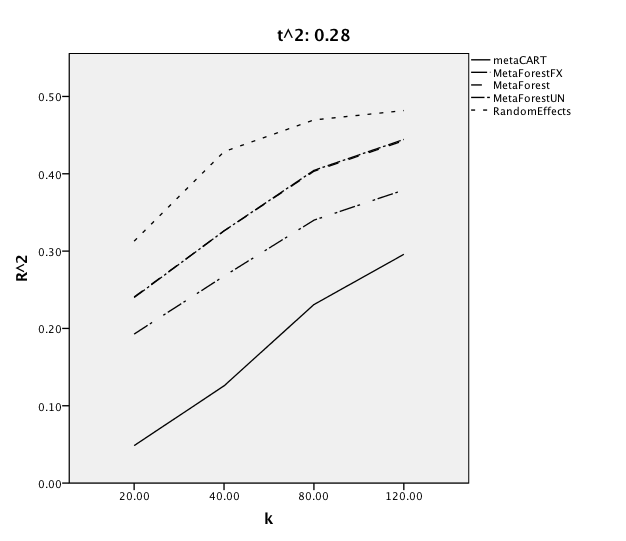
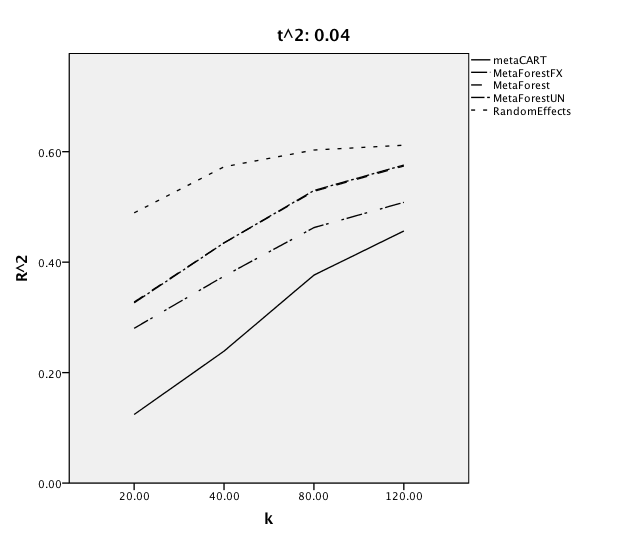




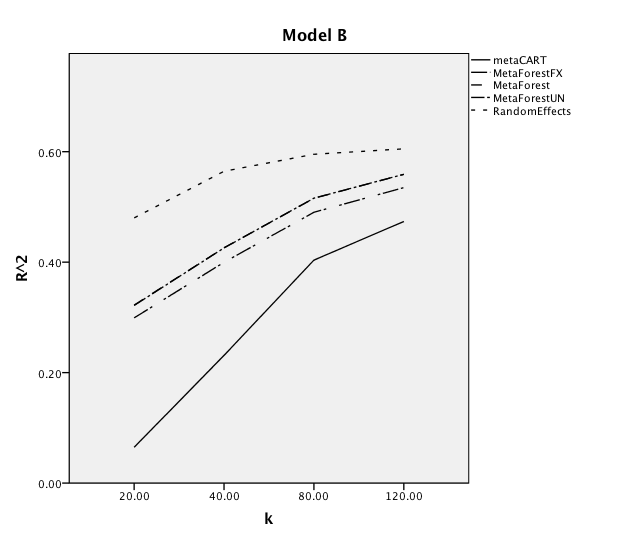
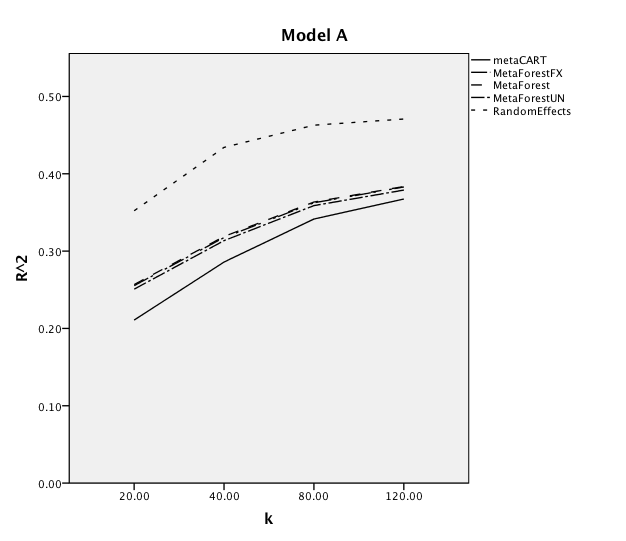


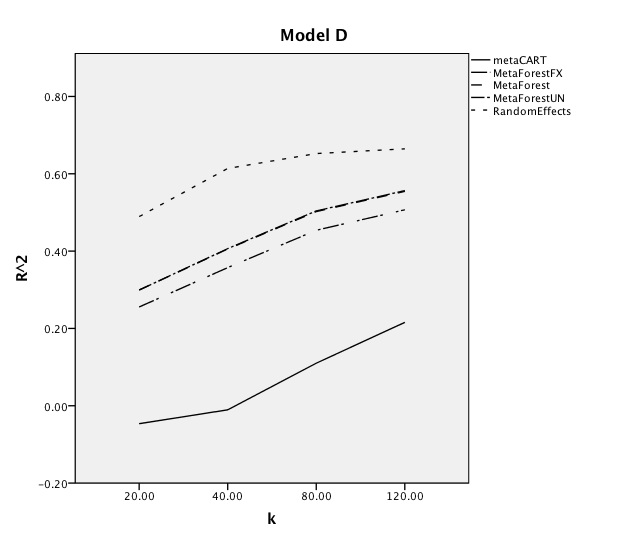
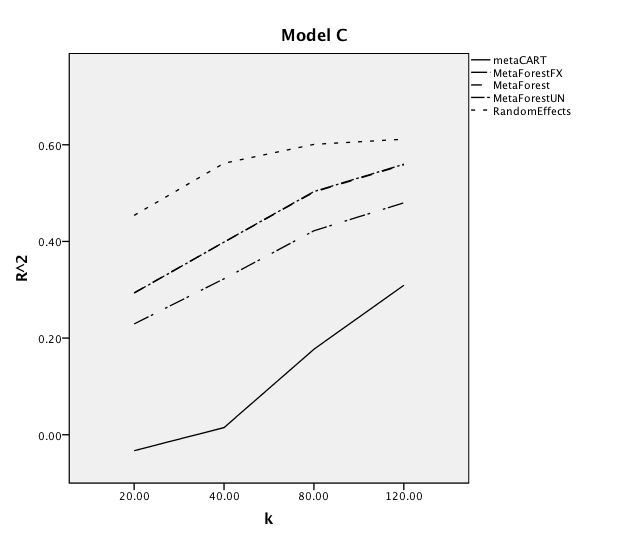
*Figure 3.* Marginal for the interaction between the number of studies and residual heterogeneity

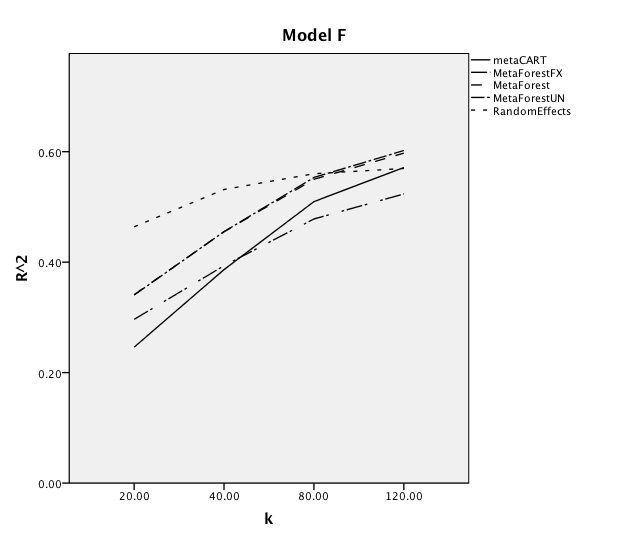
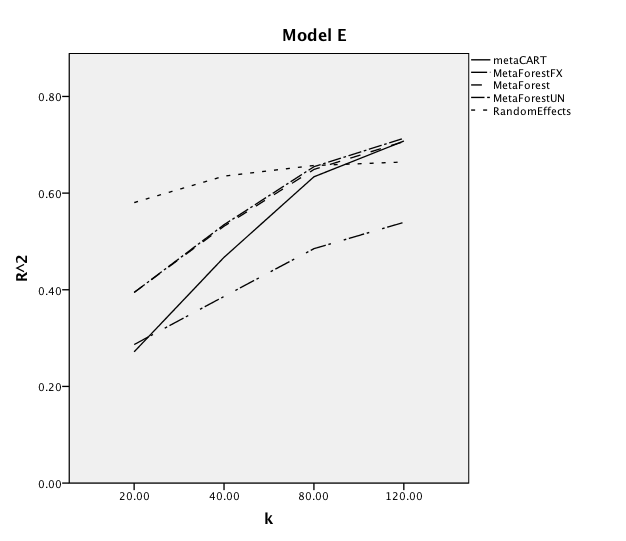




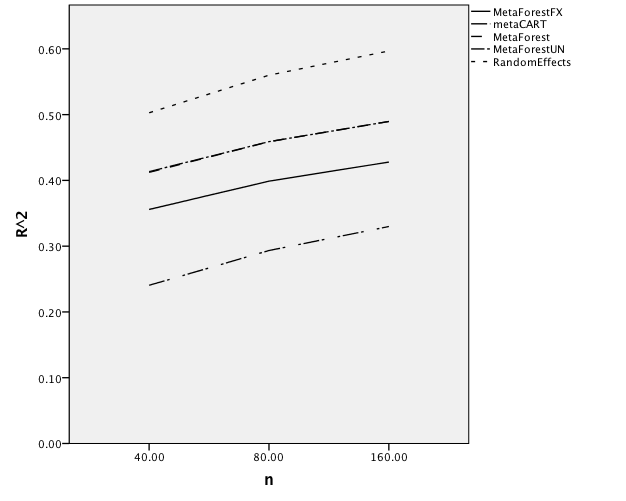
*Figure 4.* Marginal for the interaction between the number of studies and the model



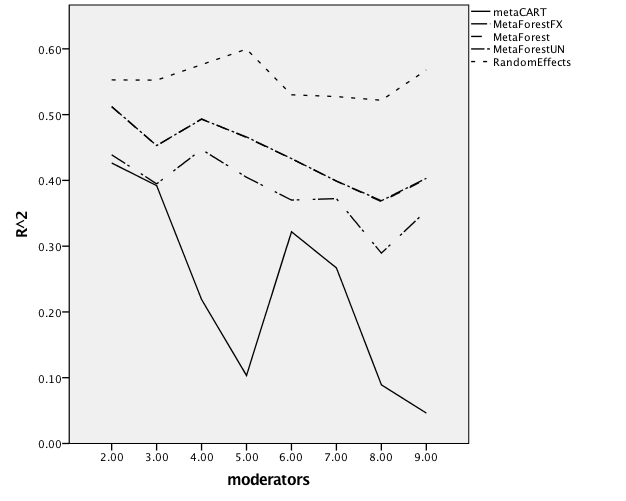




*Figure 5*. Marginal for the effect of the average within-study sample size



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



**References**

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

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.

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

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