Bachelor Thesis

Exploring the use of a lasso algorithm in a moderator problem in meta-analysis

Date of submission: 19-7-2019

Student: A.J.C. (Antoon) van Beek (5981689)

Supervisor: C.J. van Lissa

Methods & Statistics department Utrecht University

Introduction

In meta-analysis, it is often the case that a lot of between study variance is present, but the exact sources remain unexplored. An exploratory moderator analysis investigates the study heterogeneity by including study characteristics as covariates. However, when performing such an analysis, a lot of characteristics, or "moderators", are measured and it may be unclear which are relevant and which are not. This, together with the fact that in meta-analysis the number of included studies can be fairly low, causes some trouble when trying to correctly perform a meta-analysis. This trouble has to do with the interpretability and the predictive power of the model that gets fitted by the meta-analysis. Earlier efforts to solve the problem of moderators within a meta-analytic setting proved to perform really well under these circumstances. The tree-based meta-analytic tool proved to have sufficient power at a low amount of studies included when the moderators were continuous (Van Lissa, 2017). However, in a situation where moderators are binary, there is still much to be gained. We intend to solve this problem by making use of penalized regression. But before it seems reasonable to clarify some core principles about meta-analysis and its importance in research in general.

In recent years, the need for making conflicting and complicated results across studies more accessible and usable has increased. A good way to do this is by undertaking a systematic review of the existing literature of that topic (Bambra, 2011). Primary to new research, usually a literature review is performed of the existing literature on the same topic. A systematic review is more or less the same as an ordinary literature review, but it tries to do it in a thorough and fair way (Kitchenham, 2004). The thoroughness is reflected in that a systematic review does not only provide a simple overview of the literature, but it tries to identify, select, synthesize and value all research evidence relevant to a certain topic (Neely et al., 2010). Often an integral part of systematic reviewing is the review quantitative data of the individual studies. The most common and effective way to do this is with meta-analysis.

Meta-analysis is a statistical method which utilizes several tools to synthesize the data of multiple studies on the same topic, with the purpose of finding a result that is more trustworthy. What meta-analysis does is simply weighting all the observed effect sizes of the individual studies and averaging them to one summary effect. Although this explanation is a bit too simplistic, in essence this is what meta-analysis is about. Meta-analysis assigns weights to each individual study based on different assumptions which are set in advance. These weights determine to what extent an individual study takes part in the eventual summary effect.

Classic meta-analytic approaches

The two classic approaches of meta-analysis refer to fundamental different assumptions made about the underlying data. These assumptions define the weights and will also determine which methods are used for the weighting of individual studies and for the creation a summary effect.

The first approach is referred to as the as the fixed-effect model. This model assumes that each observed effect size, obtained from each individual study, is an estimate of an underlying true effect size (Hedges & Vevea, 1998). The true effect sizes are treated as, unknown, constants. The only source that causes the deviation of the observed effect from the, unknown, true effects is sampling error. Thus, for a collection of k studies, the observed effects size y_i *of* each individual study i (for i = 1,2, . . . k) is given by:

$$y_i = \theta + \epsilon_i \quad (1)$$

Where θ is the true effect size of each individual study i and ϵ_i follows the distribution of $N(0,v_i)$ with v_i being the sampling error or within-study variance, which is treated as a known factor.

The second model is the random-effects model. This approach makes an additional assumption, namely about the true effect sizes. Where the fixed-effect model treats the true effects as constants, the random-effect model assumes that the true effects are random and follow a distribution of their own (Hedges & Vevea, 1998). This means that variation in the observed effects (y_i) in the random model incorporates not only the sampling error but also the variation of the true effect sizes (τ^2) between the studies. In the case of the random effect model the observed effect size of y_i is, given by:

$$y_i = \theta_i + \epsilon_i \quad (2)$$

With $\epsilon_i \sim N(0,v_i)$ but, in this case θ_i on itself is given by:

$$\theta_i = \mu + \zeta_i \quad (3)$$

With μ being the mean of the distribution of the true effect sizes and ζ_i following the distribution N(0,τ2) with τ2 being the variance of the population of true effect sizes. It could also be explained as the variance between the individual studies.

In both models we are interested in the summary effect of all the individual studies. In the case of the fixed-effect model it is natural to estimate this summary effect by pooling from all the individual observed effects. However, individual studies with a low sampling error possess the ability to estimate the underlying true effect more accurately, thus it could be argued that the studies with a lower sampling error should weight more in the eventual summary effect. This means that a lower error variance should lead to a higher weight. The individual weights (W_i) in the fixed-effect model are given by:

$$W_i = \frac{1}{v_i} \quad (4)$$

The assumptions that sampling error is the only source of variation, makes it in this case the only factor which is important for the process of assessing weights to each study. The assumptions that sampling error is the only source of variation, makes it in this case the only factor which is important for the process of assessing weights to each study. However, in the case of random-effects, the true effects also follow a distribution, so therefore the between study variance is also taken into account when composing the weights for the individual studies. The individual weights for the random-effect model are given by:

$$W_i = \frac{1}{v_i + \hat{\tau}^2} \quad (5)$$

In the case of the random-effect model, the within-study- and between-study variance is necessary for the calculation of the weights. It is important to note that in the calculation of the individual weights, an estimation of study heterogeneity is used ($\hat{\tau}^2$). While the sampling error is known for each individual study, the true effect heterogeneity (τ^2) remains unknown. Therefore, an estimation of the heterogeneity value needs to be made to effectively calculate the weights. This estimation of the between-study variance is thus represented by $\hat{\tau}^2$.

Meta-regression

While these two models are presented here as two possible approaches to meta-analysis, it seems that the assumption of a fixed-effect seems to rarely hold in social sciences. The pursuit of capturing human behaviour in research remains very complex (Earp & Trafimow, 2015). It has been shown that conducting a perfect replication of a study of social sciences is just about impossible. The main reason for this, is that similar research questions are studied in different laboratories, using different methods, instruments and samples (Van Lissa, 2017). This may cause substantial between-study heterogeneity and this does influences how interpretable the conclusions of the meta-analysis are (Higgins & Thompson, 2002). However, simply recognizing heterogeneity is not enough. The meaning and source of the heterogeneity should be explored (Baker et al., 2009; Higgins, Thompson & Spiegelhalter, 2009). The characteristics on which studies of the same topic may differ, better known as “moderators”, could explain some of the heterogeneity in the effect sizes and should therefore be investigated. The process of examining the relationship between study characteristics and the effect sizes is most often done by a meta-regression (Viechtbauer & López-López, 2015). Meta-regression aims to relate the size of the effect to one or more characteristics of the studies involved. As multiple regression is used to assess the relationship between subject-level covariates and an outcome, meta-regression in meta-analysis is used to assess the relationship between study-level covariates and the effect size. In the case of fixed- and random-effect meta-analysis, the observed

effects are treated as estimations of the underlying true effect. In meta-regression the observed effects are estimated by the including the moderators. In other words, the true effect is now replaced by the moderator effects. This is expressed with the following equation, where θ_i represents the underlying true effect, *x* the moderators, the coefficients, with *p* being the number of moderators:

$$\theta_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \zeta_i \quad (6)$$

When this is substituted in the original equation it will result in:

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p + \zeta_i + \epsilon_i \quad (7)$$

The error term ζ_i captures the residual heterogeneity after accounting for the moderators. This term is still included because it is often the case that there still remains heterogeneity unexplained after accounting for the moderators (Thompson & Sharp, 1999). In this model the moderator effects are treated as fixed and the residual heterogeneity as random. Therefore, it is referred to as a mixed-effect meta-regression analysis model, in short, ME-MRA (Viechtbauer & López-López, 2015). To solve this ME-MRA model, both the residual heterogeneity and the moderator coefficients need to be estimated. An accurate estimation of the residual heterogeneity contributes to a better interpretation of the effect of the moderators (Panityakul, Bumrungsup & Knapp, 2013).

Estimating residual heterogeneity

The topic of estimating the residual heterogeneity is a highly discussed one (Veroniki et al., 2016; Viechtbauer & López-López, 2015; Panityakul et al., 2013). Numerous methods have been proposed to accurately estimate the residual heterogeneity, including the Hedges (HE), DerSimonian–Laird/Method of Moments (DL), Sidik and Jonkman (SJ), Maximum Likelihood (ML), Restricted Maximum Likelihood (REML), and Empirical Bayes (EB) method. These methods are mostly divided into two groups: closed-form or non-iterative methods and iterative methods. The main difference between these groups is that the closed form group uses a predetermined number of steps to provide an estimation for the residual heterogeneity, whereas the iterative methods run multiple iteration, as the name suggests, to converge to a solution when a specific criterion is met. It is important to note that some iterative methods do not produce a solution when they fail to converge after a predetermined amount of iteration.

The ability of the estimators to predict the residual heterogeneity is influenced by different factors, such as the number of studies (Guolo & Varin, 2017; Panityakul et al., 2013; Hardy & Thompson, 1996) included and the sample size of the individual studies (Panityakul et al., 2013). In our scenario we are especially interested in an estimator which performs well under the condition of a relative low number of studies. The Restricted Maximum Likelihood (REML) seems to produce the lowest bias under this condition and is therefore preferred (Panityakul et al., 2013; Hardy &

Thompson, 1996). The REML will be used in this study for the estimation of the residual heterogeneity.

As said before, the REML is an iterative method. This iterative method needs a starting estimation of τ^2 to start, usually it gets estimated by one of the non-iterative methods (Viechtbauer & López-López, 2015). Besides the starting value of τ^2, it needs in every iteration an estimation of the regression coefficients of the moderators. These are typically estimated by using the Weighted Least Squares (WLS) method. This is a variation of the Ordinary Least Squares (OLS), but in the case of meta-analysis it is necessary to assess weights to the coefficients. In systematic reviews large variation in standard errors is often observed, which will result in large heteroscedasticity in the estimation of the effects (Stanley & Doucouliagos, 2017). The addition of weights is a way to adjust for this heteroscedasticity. The weights are formulated as presented in equation (5).

The usage of a WLS method to estimate the regression coefficient may be problematic in the situation where a lot of moderators are measured without their specific effects, when the amount of studies is low and when moderators are dichotomous. The use of a least squares method will cause problems with the *prediction accuracy* and the *model interpretability* (James, Witten, Hastie, & Tibshirani, 2013). In the situation where a lot of moderators are measured and blindly included in the model, it may as well be the case that variables are included that are in fact not associated with the response. Including irrelevant variables in the model lowers the interpretability of the model (James et al., 2013). An approach is necessary that automatically excludes the variables that are irrelevant i.e. performs variable selection. As explained before, in meta-analysis it is often the case that the number of moderators closely approaches or even exceeds the number of studies included in the analysis. A least squares method will display a lot variability in the fit when the number of variables is not much smaller than the number of studies (James et al., 2013). This means that the least squares method over fits the data and loses its power to be generalizable to future observations. When the number of variables exceeds the number of studies, the least squares method fails to produce one unique estimate and the method should not be used at all.

However, a least squares method could still be somewhat valuable in some situations. It is extremely suitable to estimate a linear relationship. In the case of dichotomous moderators, the relationship is always perfectly linear. A powerful non-linear estimation tool is in the situation of dichotomous moderators unnecessary and would not perform better at all. Whenever a non-linear relation gets fitted on data with an underlying linear relation, it will cause problems when this fit gets used for the prediction of future data. Given the various arguments, this paper provides an approach to tackle this problem of the least squares methods whilst still making use of a linear method. The weighted least squares are replaced with the so-called LASSO (least absolute shrinkage and selection operator) regression for the estimation of the regression coefficients. This algorithm shrinks or penalizes the regression coefficients and performs variable selection (James et al., 2013; Hesterberg, Choi, Meier, & Fraley, 2008).

The Lasso

The *lasso* is a technique that regularizes or constrains the coefficient estimates, better known as *shrinking* (James et al., 2013). It possesses the ability to reduce the regression coefficient even to a value of zero. By doing this it automatically performs variable selection. It does not seem to be immediately clear why shrinking the coefficients should be an improvement to the model. However, by shrinking the parameters, it lowers the variance of the model by increasing the bias only a little bit. In other words, the model sacrifices some of its ability to fit the current data, to greatly increase the ability to predict future data with the same fit (James et al., 2013). This is better known as the bias/variance tradeoff (Briscoe & Feldman, 2011).

The Lasso shrinkage method is not the only shrinkage method, there do exist some others. Nevertheless, the lasso is in the case the best option. It possesses, as opposed to other methods, the ability to shrink the parameter not towards zero, but to be exactly zero (James et al., 2013; Hesterberg, Choi, Meier, & Fraley, 2008). This means that the lasso can perform variable selection, something that is specifically aimed for in this study.

In line with other shrinkage methods the lasso makes use of a shrinkage penalty. This penalty is added in the process of the OLS calculation of the regression coefficients. The OLS method estimates the coefficients by minimizing the Residual Sum of Squares (RSS). The following equation shows how the calculation of the RSS together with the shrinkage penalty:

$$RSS = \sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j| \quad (8)$$

This equation shows that the shrinkage penalty consists of two variables, the tuning parameter lambda (λ) and the regression coefficients (β). This means that, while the OLS tries to find the coefficients which explain as much variance as possible, due to the minimization of the RSS, the shrinkage penalty punishes this. Therefore, the coefficients are forced to shrink a certain amount, depending on the parameter lambda. If the lambda increases, it grows the impact of the shrinkage penalty on the RSS, with $\lambda \to \infty$ shrinking all the coefficient to be zero, producing the null model. But, if the lambda is zero, the shrinkage penalty has no impact at all and it will produce the OLS estimates.

Algorithms

The goal of the present study was to test whether a ME-MRA model with the *lasso* algorithm is able to outperform the ME-MRA with least squares regression. More specifically, if the lasso is able to outperform the least squares when in situation where the amount studies included in the

analysis is fairly low. To test this, two different algorithms are used; one called the *rma*, which makes use of the WLS regression, and the *lma*, which makes use of a penalized lasso regression.

The *rma* algorithm is part of the software-package **metafor** in R, which is developed by Wolfgang Viechtbauer (2010, 2019). This algorithm is specifically developed to perform a meta-analysis or met-regression. It allows to include different models, such as the fixed-, random- and mixed-effect model. It is also possible to account for moderators (Viechtbauer, 2010). The mixed-effect model, which is used is this study, requires a two-step approach to fit a meta-analytic model. First the residual heterogeneity is estimated. The package developed by Viechtbauer does provide multiple methods for the estimation of the residual heterogeneity. In this study the Restricted Maximum-likelihood is used, but this has already been discussed earlier. The second step is estimating the moderator coefficients, which is done by using the Weighted Least Squares (WLS) method. The weights are described in equation (5). The *lma* is a variation of the *rma* algorithm which is created by Caspar van Lissa. As explained before, the REML is an iterative procedure for the estimation of the residual heterogeneity. In every step of the process, instead of estimating the coefficients of the moderators by using a WLS, a weighted lasso regression is performed. Then again, the residual heterogeneity gets estimated with the rma algorithm by using the new values of the coefficients. With these new values of τ^2, a new weighted lasso is performed for the estimations of the coefficients. This process continuous, until the residual heterogeneity converges to a certain value. The algorithms are evaluated on three different performance criteria: The algorithms' predictive performance, their ability to estimate the residual heterogeneity and their ability to detect and select the right moderators.

Performance criteria

The predictive performance of the algorithms is defined by how well the algorithm is able to predict future data. The algorithms have to estimate a model on a "training" dataset and then use this model to see how well it fits on a second "testing" dataset. This is operationalized as the cross-validated R^2_{cv} (Van Lissa, 2017). The R^2_{cv} is calculated using the fraction of variance explained by the model on the testing dataset, relative to faction of variance explained by the mean of the testing dataset. The mean of the testing dataset is the best prediction for the testing data when there is no model present (van Lissa, 2017). The calculation of R^2_{cv} is expressed by the following equation:

$$R^2_{cv} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \quad (9)$$

With *n* being the number of studies in the testing dataset, $\hat{y}_i$ being the estimation for study *i*, and $\bar{y}$ being the mean of the training dataset.

The ability of the algorithms to estimate the residual heterogeneity is by simply taking the value of τ^2 which to algorithm produces. The true value of the residual heterogeneity is subtracted of

the estimated value, solely to make the values more interpretable. This means that a correct estimation of the residual heterogeneity will be expressed by a value which exactly or close to zero. The residual heterogeneity is used as a performance criterion because it is suspected that the lma model might not always be able to predict residual heterogeneity correctly.

The ability of the algorithms to detect and select the right moderators is defined by looking at the fractions of true- positives and negatives. This is further operationalized by taking the product of those relative to the sum of fractions of true- positives and negatives[1]. This is done because evaluation the fraction of the negatives and positives individually will not provide a good inside. A badly fitted model can still have a perfect score on the number of true negatives it detects, while having detected none of the true positives. This operationalization punishes a poorly performed detection of either the positives or the negatives heavily, so that bad models will have a low value on this criterion. The calculation of this performance criteria, further referred to as *FPS* (Fraction of the Product relative to the Sum), is expressed in the following equation:

$$FPS = \frac{(P_i * N_i)}{(P_i + N_i)} * 2 \tag{10}$$

With P_i being the fraction of positives for study *i* and N_i being the fraction of negatives for study *i*.

To test the lma and rma algorithms on the performance criteria, a simulation study is performed. A simulation of the data is preferred over the use of real data. Simulated data can be shaped to such an extent that it will have the all desired characteristics to test the performance of the algorithm. Besides that, if simulated correctly, it will not have any systematic errors or noise due to underlying models and it is more cost efficient.

In the simulation study, meta analytic datasets will be simulated. These datasets consist of two separate sub-datasets, a training- and a testing dataset. Both sub-datasets will have the same characteristics with the exception of the number of studies included. Certain characteristics of the sub-datasets will be manipulated to test how well the algorithms perform under certain conditions. For each combination of characteristics, or design factors, 100 datasets will be simulated. The design factors that will be manipulated are the number of studies in the training data k (22, 40 and 80), the average within-study sample size $\bar{n}$ (40, 100 and 200), the population effect size β (.2, .5 and .8) and the residual heterogeneity τ^2 (.01, .04 and .1). All the datasets will contain 20 moderators of which 10 are relevant and 10 are irrelevant. The moderators are binary and are randomly drawn form a Bernoulli distribution with probability $p = .5$, which corresponds to an equal chance of being either one or zero. The dependent variable y_i represented by a Hedges' g. This is an estimator which takes the standardized mean difference between a treatment and control group and is commonly used in meta-analysis (Van Lissa, 2017). The true effect size θ_i is sampled out of a normal distribution. The

[1] *I was not able to find a source which validates this operationalization. However, this does not imply that I was the first to come up with this operationalization.*

mean is computed by the assessing the values of the coefficients β, with the values of the moderators and with the residual heterogeneity τ^2 (Van Lissa, 2017). This is in line with the calculation of θ_i represented in equation (6). The sampling error v_i is formed by varying the sizes of the samples of each study. The sample sizes n_i are drawn from a normal distribution with mean $\bar{n}$ and standard deviation $\bar{n}/3$ (Van Lissa, 2017).

Design factors & simulation

These design factors are chosen on purpose, because they are hypothesized to have an influence on the predictive performance of the algorithms. The effect of the design factors ought to be either positive or negative on the data. This means that some factor should, by increasing, make the data easier to be analyzed, or make it more difficult to analyze. The amount of studies included in the training data k has a positive influence on the variance explained by the different algorithms. This is due to the fact that there are simply more data points available to fit a model on. The lma algorithm should be superior on the low value of k over the rma algorithm. The effect size β has a positive impact on the ability of the algorithms to explain variance. It can be hypothesized that the lma performs better at lower values of β because it is better equipped to detect and select variables when even when the amount of signal is low. The residual heterogeneity τ^2 should have a negative influence on the interpretability of the data. Differences between the two algorithms could be present, but it remains unclear which would perform better. The lma might perform better when the amount of signal in the data is low or the noise is high, but it is also suspected to overestimate the amount of heterogeneity and this could worsen if the τ^2 increases. The $\bar{n}$ greatly influences the quality of the data. Higher values of within-study sample sizes reduce the sampling error. This will lead to a better prediction by the algorithms. In conclusion: higher values of k, β and $\bar{n}$ will increase the quality of the data, where higher values of τ^2 decrease the quality of the data. The lma is suspected to perform significantly better when the quality of the data is low, especially when the amount of studies in the sample is low, with the exception of the performance of the lma on the estimation of the residual heterogeneity.

Data

Before performing the analyses, the values the algorithms produce on the performance criteria, or the dependent variables in this case, were checked. Summary- or descriptive statistics and density plots of all the dependent variables were looked at to see if there were any abnormalities (See table 1 and figure 1 & 2). Sadly, the density plots of the R^2_{cv} failed to be interpretable at all, due to the huge negative value produced by the rma algorithm (table 1). The statistics and the plots of the other

criteria did provide a good insight in the distribution of the data. Table 1 shows that there is a quite a big difference in the mean of the algorithms on $\Delta\tau^2$ (.975 for lma opposed to .0303 for the rma). In figure 2 can be seen that $\Delta\tau^2$ shows a very flat distribution for the lma algorithm, whereas the rma form around 0.

The large negative values for R_{cv}^2 in the rma algorithm are a bit problematic. These value so extremely high that they will influence the interpretability of any analysis. The high values cause the mean and standard deviation to be really absurd. A mean of -9.08 means that the rma algorithm, on average, explains nine times less variance than the null model. A real outlier analysis was not performed, because especially outliers provide a lot of insight in the performance of the algorithms. These values do need to be handled in some sort of way and therefore the possible causes of the extremes were explored. The first step in doing this is by creating a range. Any value outside this range will be marked as an extreme value. The range was based on the 1.5 times the value that marked the lowest 10% of R_{cv}^2 in the rma. This resulted in a value of -7.426. Subsequently a subset was created of the data, which only contained the cases with extreme values. All of the cases with extreme values where produced by the rma algorithm. Table 2 show the distribution of the extreme values over the other design factors. The outlier cases are more or less equally distributed over all the subgroups of the factor, with the exception of design factor k. All of the extreme values are in the subgroup $k = 22$, which provide evidence to suspect that the low value of k caused the extreme values of R_{cv}^2 in the rma. Any analyses of the effects of other design factor on R_{cv}^2 will be suppressed by the huge impact of k. For this reason, the subgroup $k = 22$ is removed and further analyses on R_{cv}^2 will be done with the subset only including $k = 40$ & 80. Table 3 shows the descriptive statistics of both subgroups and figure 3 shows the distribution of subgroup $k = 40$ & 80. This table shows that, when the subgroup is removed, the means of both algorithms are almost equal. There does exist more variance around the mean for the rma however. This is also displayed in the density plot (figure 3). The rma still has a tail that goes below a R_{cv}^2 value of 0. The lma forms a second peak around the 0 and has very few values below 0.

Analyses

For the evaluation of the effects on the dependent variables, ANOVAs where performed. In ANOVAs the main effect of the algorithms, the design factors and the interaction between those where assessed. However, due to the large sample size (16200 in total), the effects where almost always significant under a 95% confidence interval. This mean that small, rather trivial, effects will be detected as significant. Thus, it was set stricter, to a confidence interval of 99.9% interval. The intention was that a post-hoc test would be performed to look at differences between the individual groups of the design factor and the algorithms, but the Tukey's honest significance test detected very

small differences as significant, even with a confidence interval of 99,9%. Therefore, the decision was made to not include this statistical test, but to assess the means and standard deviation of each subgroup and interaction plots in a non-statistical manner. The main reason for this is to provide a more sound and broader explanation of the performance of the algorithms under certain circumstances, instead of only showing the effects and the significant levels. The evaluation of the effects will be done as follows: if the ANOVA show no significant result, either on the main effects or the interaction, then the effect will not be further discussed. If some are significant, which is most cases, the means will be compared in a subjective way. If this provides evidence for a strong effect or interaction, the interaction plot will be used to further visualize the effect.

Results

The section that follows contains the results of the analysis. Table 4 tells whether or not certain effect where significant or not. Table 5 provides the means on the dependent variables for the subgroups of the design factors for each algorithm. Figures 4 to 9 contain the interaction plots which where used to visualize the interaction effects that were significant.

The predictive performance of the algorithms was measured by the R^2_{cv}. As explained before all the analyses and the interpretation of those will be done on a subset which does not contain cases with $k = 22$. Table 3 shows that both algorithms do not differ that much in the mean, which could mean that they show the same predictive performance on this subset. Table 4 provides more evidence for this, because the algorithms never have a main effect in the column of R^2_{cv}, which means that there is no significant difference between those groups. All the design factor seems to have an effect on the predictive performance of the algorithms. The effects of k and β seems to be different for the algorithms, due to the significance of the interaction effect. Higher levels of k seem to have a significant positive effect on the predictive performance of algorithms. The means of R^2_{cv} increase from .527 and .492 in $k = 40$ to .632 and .662 in $k = 80$, respectively for the lma and rma algorithm. Figure 4 shows this increase in the form of an interaction plot. Here the interaction is better visualized and it can be seen quite clear that, where the lma outperforms the rma in $k = 40$, the rma performs better at $k = 80$. The effect sizes or β also have a positive influence on the predictive performance. The means of R^2_{cv} increase from, respectively for the lma and rma algorithm, .271 and .227 when $\beta =$ .2 to .684 and .701 in $\beta = .5$ and from this to .784 and .804. The effect of β seem t flatten out on the between the highest and the middle value of β. Also, the predictive performance of the lma seems to be, relative to the rma, higher in the lowest value of β whereas the rma takes over $\beta =$ .5 & .8 (figure 5). The τ^2 and $\bar{n}$ only show a main effect and no interaction. The R^2_{cv} decreases for higher values of τ^2 and increases for higher values of $\bar{n}$. However, it is important to keep in mind that the $k = 22$ group is not included in all the analyses. If that was the case, more differences would have been

observed between the main effects and there might have been differences in the main effects of the design factors and the main effects of the interaction effects between the design factors and the algorithms.

The estimation of the residual heterogeneity was measured by the dependent variable $\Delta\tau^2$. There is always a main effect of the algorithms of the main effects, which means that there is a significant difference between the algorithm. Table 1 and 5 show that the lma always seem to over estimate the residual heterogeneity, whereas the rma seems to produce values relatively close to zero, which suggests a correct prediction of the residual heterogeneity. The design factors k and β show to have significant main effect (table 4). Table 4 also suggest that there is an interaction effect between the algorithms and the design factors β and $\bar{n}$. The value of heterogeneity in the dataset does not seem to have an impact on the estimation of it. The main effect of k seem to have a slight negative effect on the values of $\Delta\tau^2$ which suggest a positive effect on the estimation of the residual heterogeneity (table 5). However, the differences of the algorithms still seem substantial across the values of $\Delta\tau^2$ with values of .978 to .967 for the lma and values of .081 to .002 for the rma. Surprisingly, increasing effect sizes of the coefficients seem to have a negative effect on the estimation of the residual heterogeneity. The values of $\Delta\tau^2$ increase for higher values of β, especially for the lma (table 5 & figure 6). The values of $\Delta\tau^2$ seem to be relatively close for $\beta = 0.2$ compared to difference between the algorithms at $\beta = 0.8$. Such an increase or decrease in the values of $\Delta\tau^2$ is not observed for any of the other design factors. However, the effect of $\bar{n}$ is quite remarkable. The main effect of the within-study sample size gets suppressed by the interaction effect. The interaction effect shows that the effect $\bar{n}$ is in opposite direction for the lma and the rma (figure 7). The values for the lma increase from .954 at $\bar{n} = 40$ to .989 at $\bar{n} = 200$, whereas the values of the rma decrease form .063 $\bar{n} = 40$ to .006 at $\bar{n} = 200$.

The ability of the algorithms to detect and select variables was expressed by the dependent variable FPS. The ANOVAs suggest that there is a significant difference between the algorithms (table 4). The lma algorithm seem to almost always better detection of amount of relevant an irrelevant moderator. Every design factor shows to have a significant impact on this as well. The k, β and $\bar{n}$ have a significant positive effect on FPS and the τ^2 has a significant negative effect. For the design factor k and β also show to have an interaction effect with the algorithms. The effect of k show to increase to be higher for the rma algorithm. The mean of FPS increases from .509 in $k = 22$ to .736 in $k = 40$ and to .865 $k = 80$. For the rma this increase goes from .264 in $k = 22$ to .761 in $k = 40$ and to .915 in $k = 80$. Where the lma outperforms the rma on $k = 22$, the rma has a higher mean of FPS at $k = 40$ and $k = 80$. This is visualized in the interaction plot (figure 8). The effect size factor displays a higher increase of FPS for the lma on higher values of β. Figure 9 shows that the values for the lma are at every value of β higher, but that this the difference increases at higher values of β. The difference increases from $.503 - .481 = .022$ to $.821 - .755 = .066$. However,

between $\beta = .5$ and $\beta = .8$ this difference does decrease again. For τ^2 and $\bar{n}$ there is no interaction. The main effects seem to be same across the groups. The lma outperforms the rma in all the situation there.

Discussion

Overall, not taken the effects of other factor into account, the lma has a better predictive performance and is more equipped to perform variable selection. However, when estimating the residual heterogeneity, the lma is inferior to the rma. The rma does start to outperform the lma algorithm when there are sufficient studies included. The design factors all influence the performance of the algorithms, with some exceptions.

The predictive performance of the rma model is heavily influenced by the number of studies. To such an extend where the rma does not produce any results when the amount of studies is too low. The claim can be made the lma is superior over the rma when a relative low number of studies in included in the analysis. However, this does not immediately mean that the lma algorithm performs good under these circumstances. The lma model almost never produces model that have less variance than the null model, but this is because the lma produces a null model when it shrinks all the coefficients to zero. In situations where there is not much data to work with, in this a low amount of studies, the amount of variance that the rma model produces is quite large, which leads to severe overfitting in future data. The lma increases its bias to tackle this. But if this result in models which are close to the null model, does that imply that produces better result. Unarguable, the lma model is less often "wrong" than the rma model when there is not much data to work with, but this does not mean that the results it does produce are all undeniably "good".

The effect of the population effect size heavily influences how the lma estimated the amount residual heterogeneity. Higher effect sizes make it more difficult for the lma to correctly predict the amount of heterogeneity which result in a severe overestimation. This could be explained by how the lma is created, namely by the integration of a lasso algorithm in the rma algorithm, which uses REML to predict the τ^2. The REML normally works with WLS values to estimate the amount of heterogeneity that is left after accounting for all the moderators. The Lasso is a more conservative method compared to the WLS and will produce lower coefficient estimates. This will make the REML "think" that there is more heterogeneity left after it accounts for the lasso coefficients. The systematic underestimation of the coefficients of the lasso algorithm can also lead to an early convergence of the REML iterative process, which could also explain for the systematic overestimation of τ^2 of the lam compared to the rma. The increase in the overestimation for higher effect sizes, could be explained that higher effect sizes are affected more by the lasso. Equation (8) shows the equation for the lasso. The first term, which contains the amount of error is not affected by

an increasing population effect size, whereas the second term, the penalty term, is affected by this. This causes the lasso to react more on higher effect sizes, which causes in larger shrinkage of the coefficients. The question remains, whether or not it is a problem that the lma algorithm is not able to correctly predict the residual heterogeneity. If one is especially interested in the amount which is present, yes in that situation it is a problem. Still, it could be argued that the estimation of the residual heterogeneity is just used as a tool for creating a better estimation of the model. The other performance criteria do not seem to be so affected by this overestimation, because the lma stills performs relatively good on these criteria. But maybe the lma could perform even better when residual heterogeneity is predicted correctly.

In the process of the operationalization of the variables and the implementation of the analyses, were decisions made which should not remain undiscussed. For the performance criterion, the ability of the algorithm to perform variable selection or detection, the fraction of true positive and true negatives, which the algorithms produce, have been operationalized to the dependent variable FPS. This variable combines both the true positives and true negatives to one value. By operationalizing the variable this way prevents the assessment true positives and negatives individually. Any differences in the between the algorithms over the true positives and -negatives are not looked at by this operationalization. These possible differences between could still provide some information of how the algorithms work. For example, a higher true negative suggest that the model produced by the algorithms is sparser, and does not include every variable. It is expected that the lma algorithm would create a more conservative model and would more often include less variables due to the shrinkage of the coefficient.

All the analyses performed where ANOVAs. The core assumptions of the ANOVAs are, that the dependent variables follow a normal distribution and that their residuals are distributed equally around the means. Both of these assumptions are in every situation violated. This does not immediately mean that the ANOVAs do not provide any valuable information, but it is something which should be assessed. In most situation where the assumptions are severely violated, a non-parametric method is advised. However, non-parametric methods do not lend themselves for the evaluation of possible interaction effects.

The in assessment of the predictive performance the decision was made to use a subset. By doing this all the effect of the of the design factor are influenced and should therefore be interpreted differently. The rma shows to perform not so well on $k = 22$. All the analyses on the other design factor do not include this group and will all hold higher values for the rma model. Also, the performance of the lma model is not evaluated on the $k = 22$, while this being a point of interest in this study. The decision for the value $k = 22$, could also be seen as not a really good one. A k with a slightly higher value could have produced better result of the rma model. Result which did not contain any extreme values and which where suitable for analysis.

Conclusion

To conclude this paper, some suggestions are made for further development of the algorithm and future research. To start, the use of lasso algorithm in meta-analysis should be further investigated, especially in the situation where there is a low amount of studies included. The ability of the algorithms to select variables was evaluated, however this could be done more thoroughly. For example, more information could be provided how the algorithms in the under- and over selection of variables. Doing this creates a better picture of what could be further developed in the lma model. Also, the estimation of residual heterogeneity remains problem of the lam algorithms which should be improved. Although, the implication of the overestimation of the residual heterogeneity on the predictive performance of the algorithm remain unclear.

Table 1: Descriptive statistics of the variables for the lma and rma (N = 8100)

		Minimum	Maximum	M	S.D.
R^2_{cv}	lma	-4.38	.963	.478	.334
	rma	-37751.	.964	-9.08	422.
$\Delta\tau^2$	lma	-.0625	5.03	.975	.849
	rma	-.1	6.80	.0303	.228
FPS	lma	0	1	.703	.313
	rma	0	1	.647	.356

Table 2: Number rma cases in the with extreme values for R^2_{cv} for each design factor (per design factor: N = 610)

Design factors	Values of the design factor	Number of cases in each group
k	22	610
	40	0
	80	0
β	0.2	359
	0.5	135
	0.8	116
τ^2	0.01	176
	0.04	192
	0.1	242
$\bar{n}$	40	261
	100	196
	200	153

Table 3: Descriptive statistics of R^2_{cv} for the lma and rma, for $k = 22$ (N = 2700) and $k = 40$ & 80 (N = 5400)

		Minimum	Maximum	M	S.D.
R^2_{cv} (22)	lma	-4.38	.938	.275	.335
	rma	-37751.	.913	-28.4	731.
R^2_{cv} (40&80)	lma	-.150	.963	.579	.284
	rma	-2.69	.964	.577	.344

Figure 1: Density plot of $\Delta\tau^2$ for the two algorithms

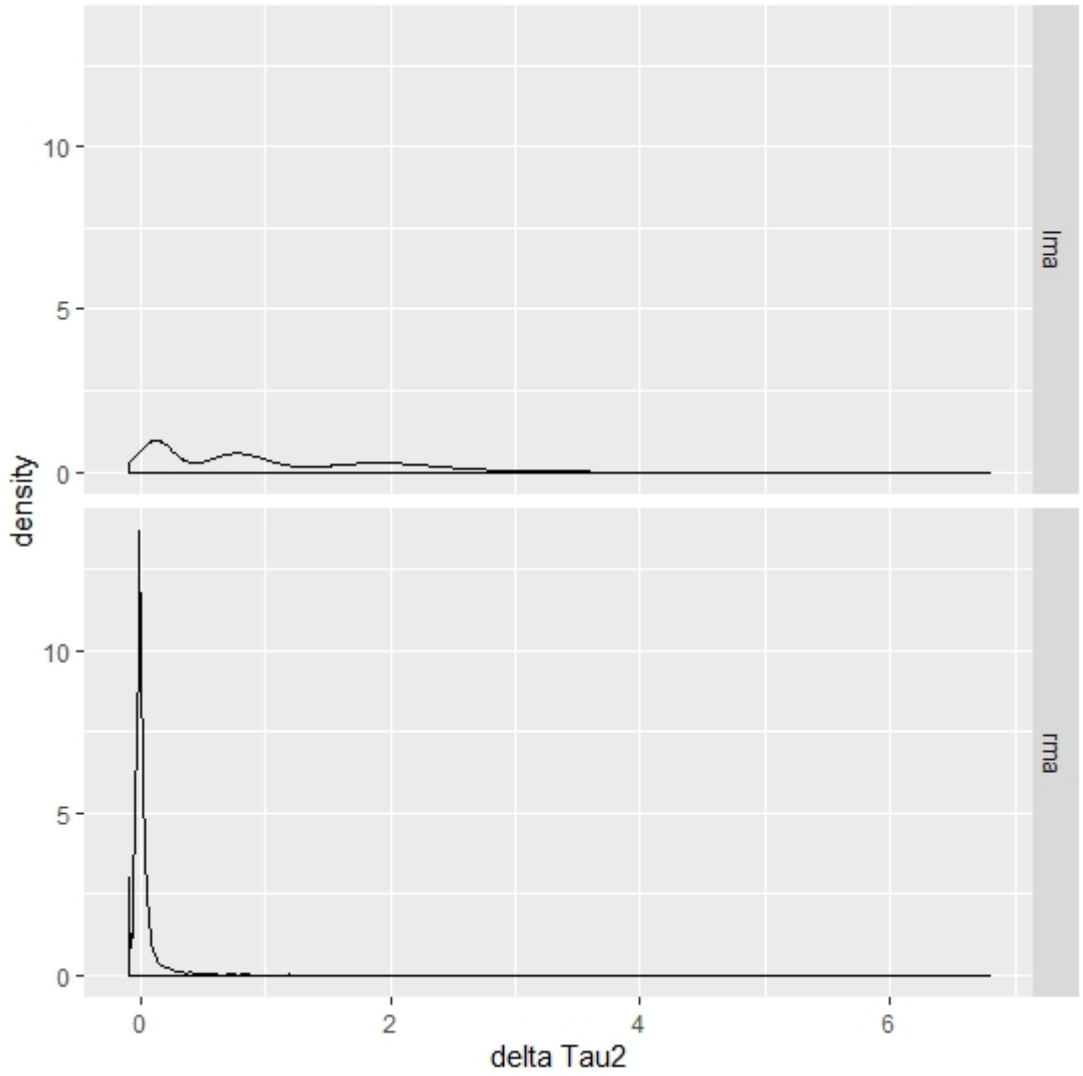


Figure 2: Density plot of *FPS* for the two algorithms

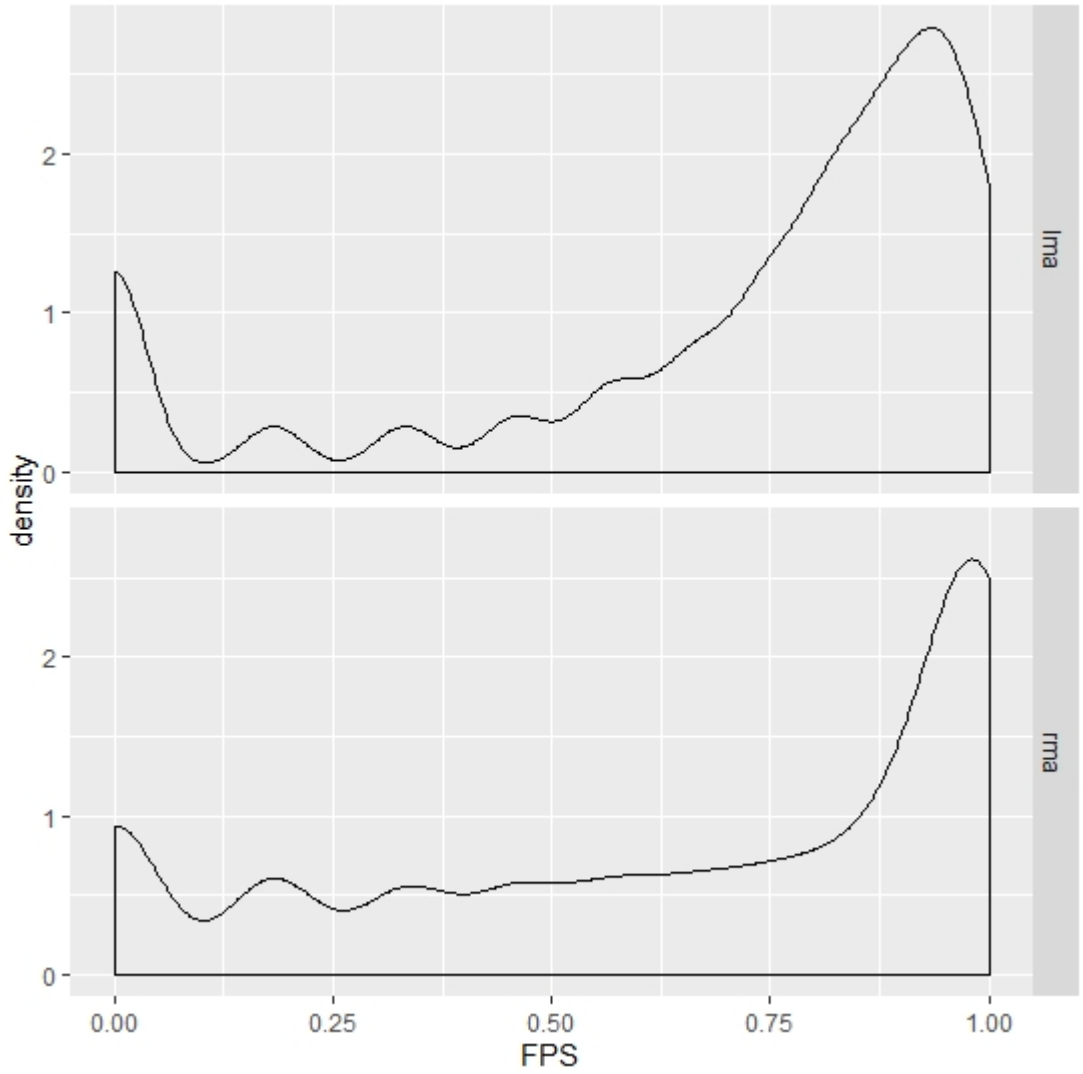


Figure 3 : Density plot of R^2_{cv} for the two algorithms on the subset $k = 40$ & 80

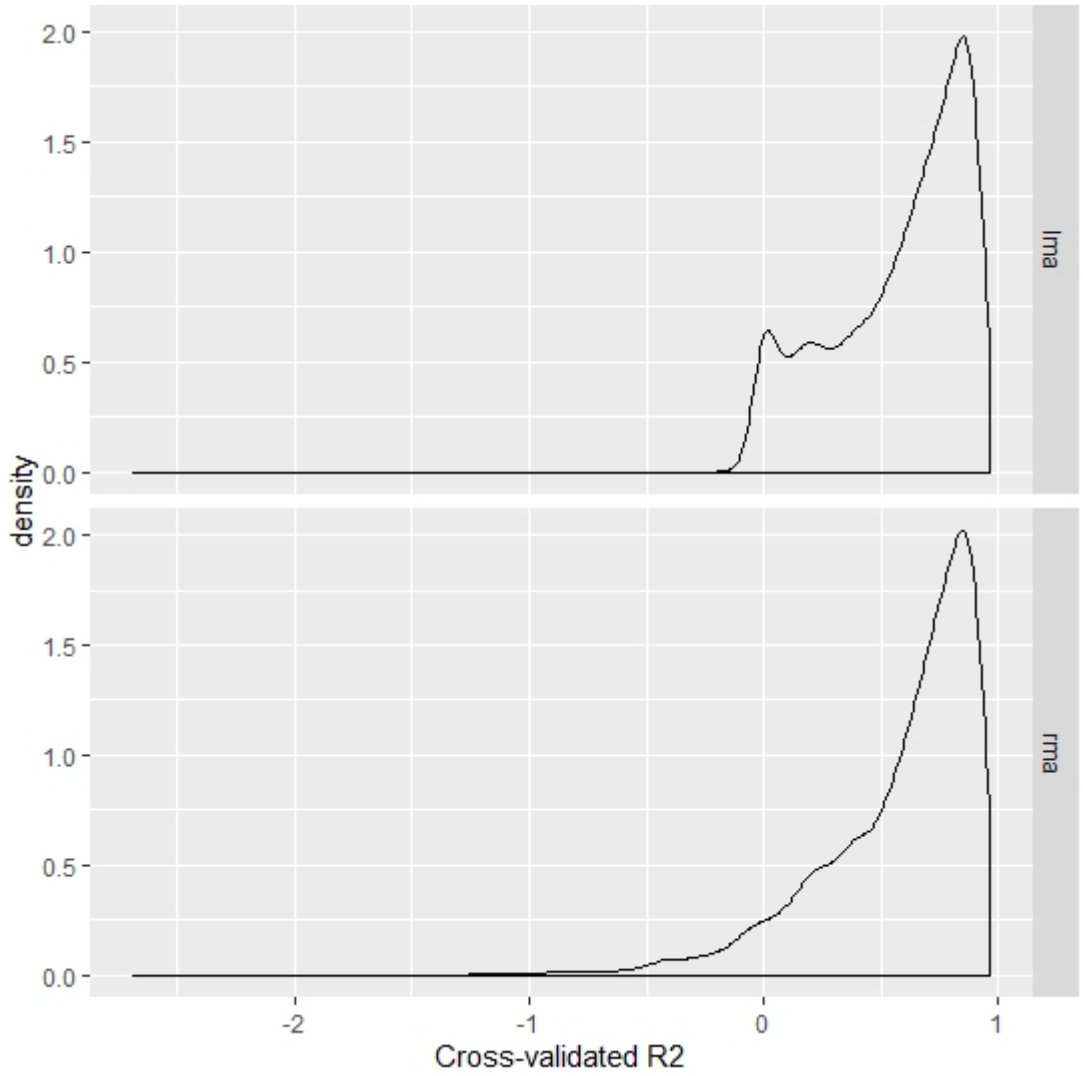


Table 4: Significance of the algorithm-, main- and interaction effects for the ANOVAs

		R^2_{cv}	$\Delta\tau^2$	FPS
k	ALG		**X**	**X**
	MAIN	**X**	**X**	**X**
	INTERACT	**X**		**X**
β	ALG		**X**	**X**
	MAIN	**X**	**X**	**X**
	INTERACT	**X**	**X**	**X**
τ^2	ALG		**X**	**X**
	MAIN	**X**		**X**
	INTERACT			
$\bar{n}$	ALG		**X**	**X**
	MAIN	**X**		**X**
	INTERACT		**X**	

Effects are significant when p < .001, ALG = Algorithm effect, MAIN = main effect of the design factor, INTERACT = interaction effect between the design factor and the algorithm

Table 5: Means of the dependent variables on the subgroup of the design factors for each algorithm

		R^2_{cv}		$\Delta\tau^2$		FPS	
		lma	rma	lma	rma	lma	rma
k	22	-	-	.976	.081	.509	.264
	40	.527	.492	.974	.007	.736	.761
	80	.632	.662	.964	.002	.865	.915
β	.2	.271	.227	.125	.012	.503	.481
	.5	.684	.701	.792	.030	.786	.704
	.8	.784	.804	2.01	.048	.821	.755
τ^2	.01	.651	.658	.978	.034	.754	.675
	.04	.591	.594	.978	.032	.712	.653
	.1	.496	.480	.967	.025	.644	.611
$\bar{n}$	40	.420	.430	.954	.063	.619	.560
	100	.611	.613	.981	.022	.725	.659
	200	.698	.699	.989	.006	.766	.721

All means are based on 2700 observation with the exception of the β, τ^2 & $\bar{n}$ under the R^2_{cv} column (1800 observation)

Figure 4: Interaction plot of the number of studies and the algorithms on R^2_{cv}

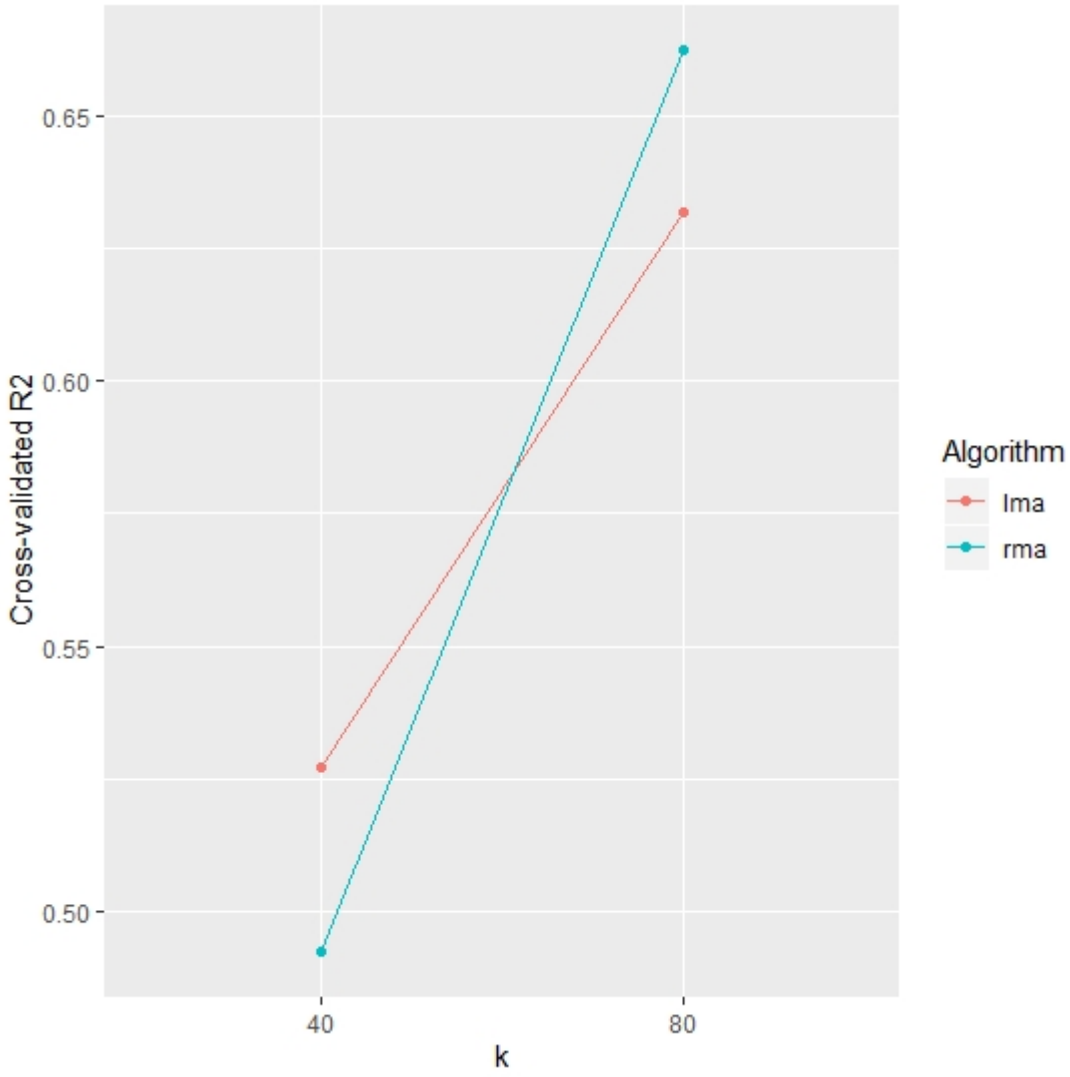
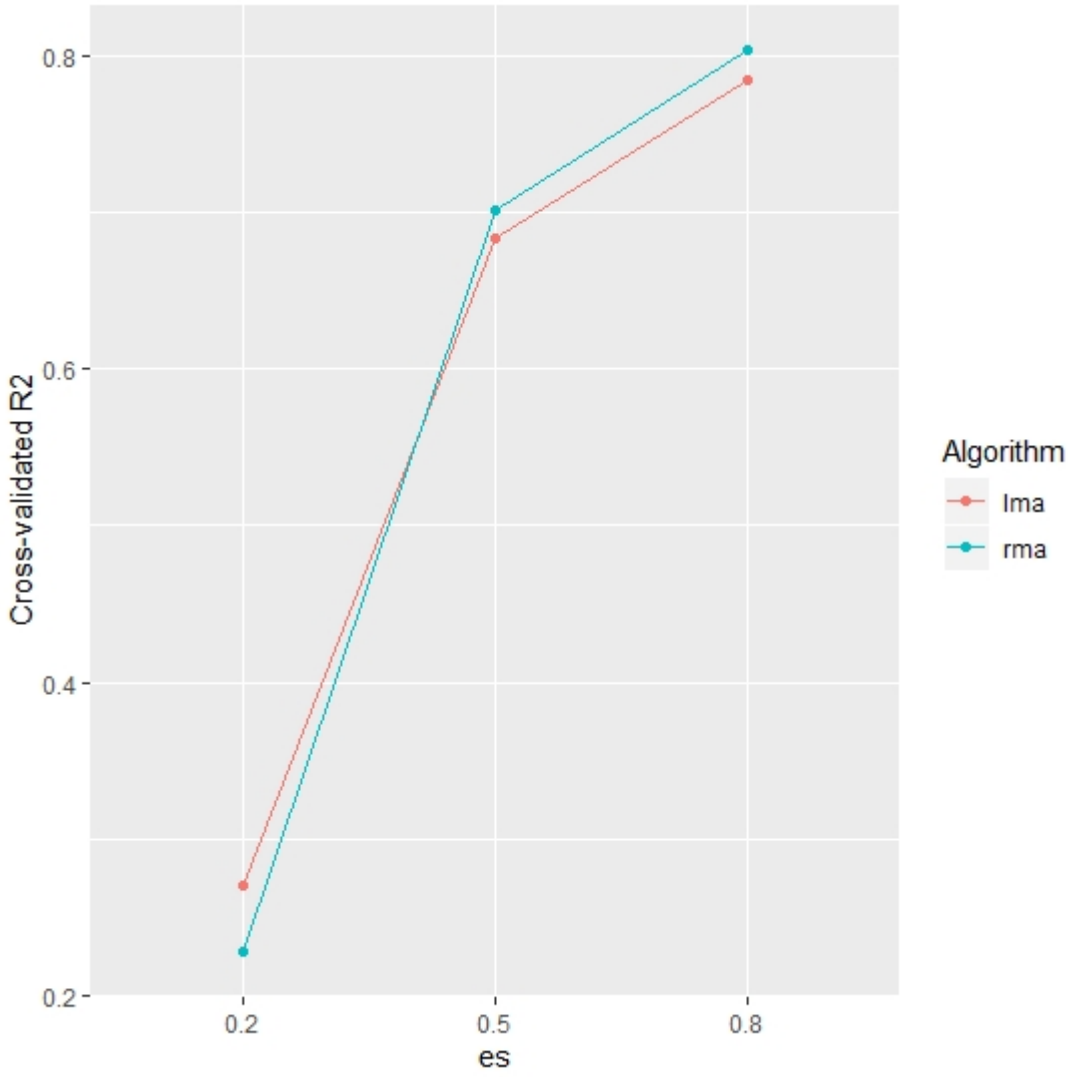


Figure 5: Interaction plot of the effect sizes and the algorithms on R^2_{cv}



This plot shows the means on the subset of the data

Figure 6: Interaction plot of the effect sizes and the algorithms on $\Delta\tau^2$

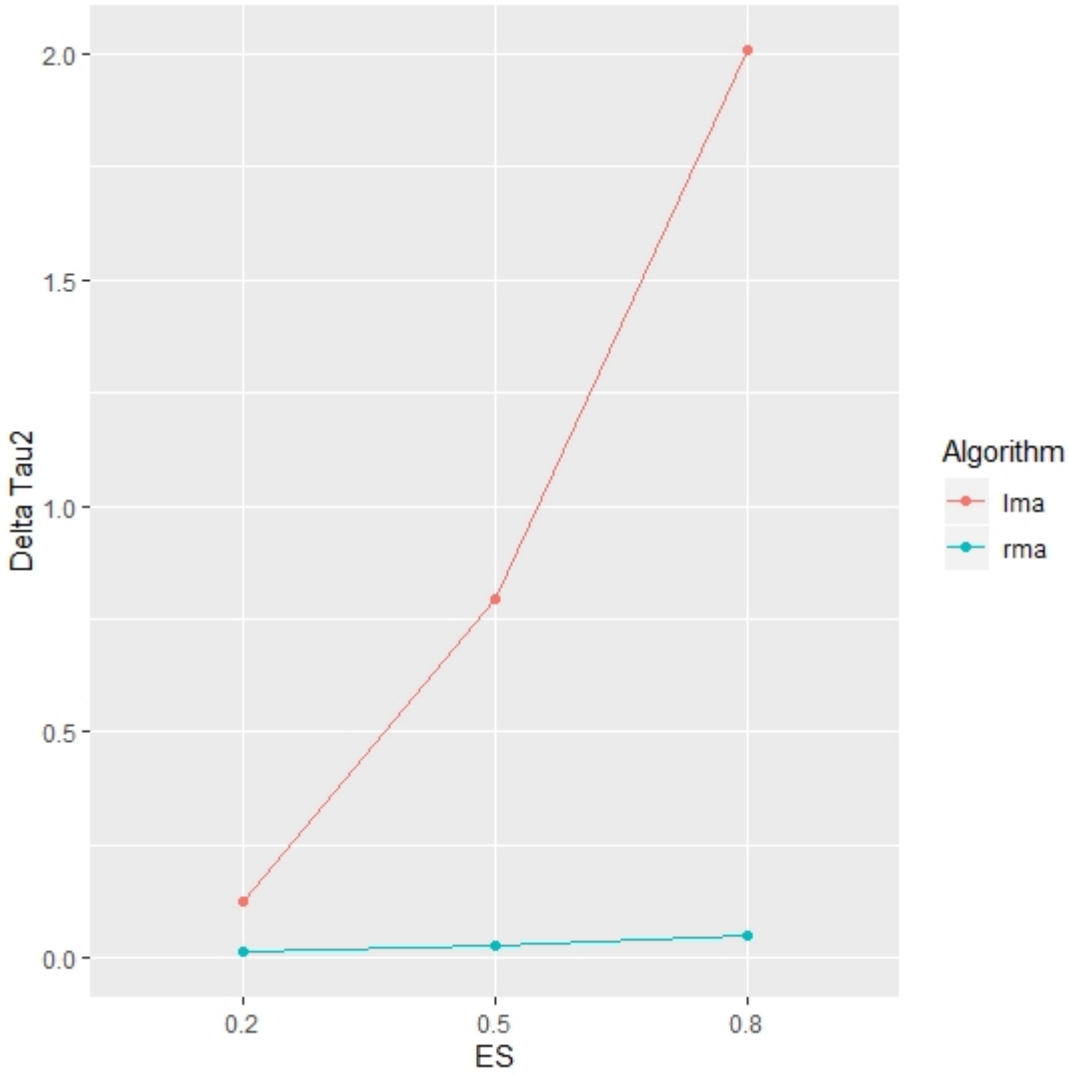


Figure 7: Interaction plot of the $\bar{n}$ and the algorithms on $\Delta\tau^2$

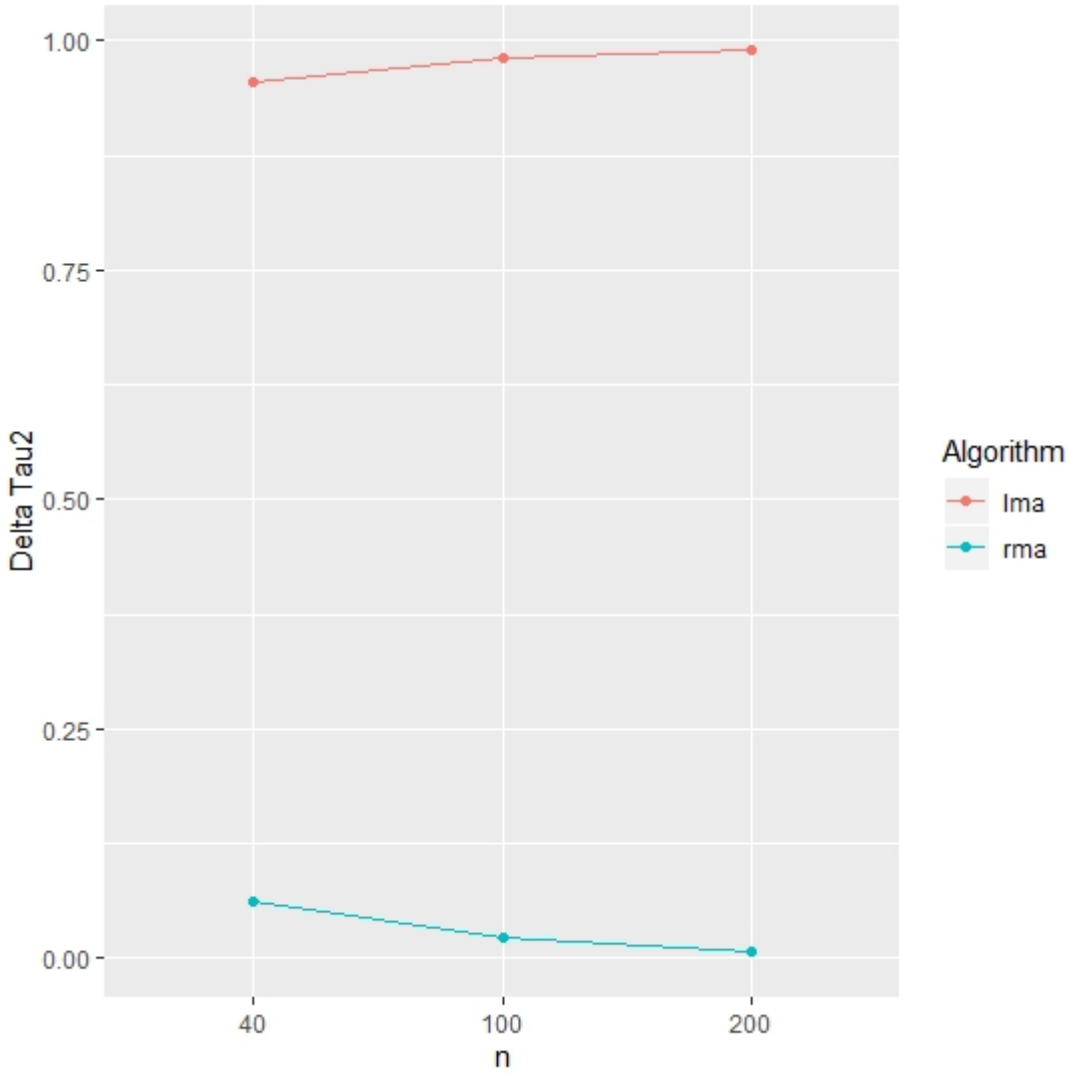


Figure 8: Interaction plot of the number of studies and the algorithms on FPS

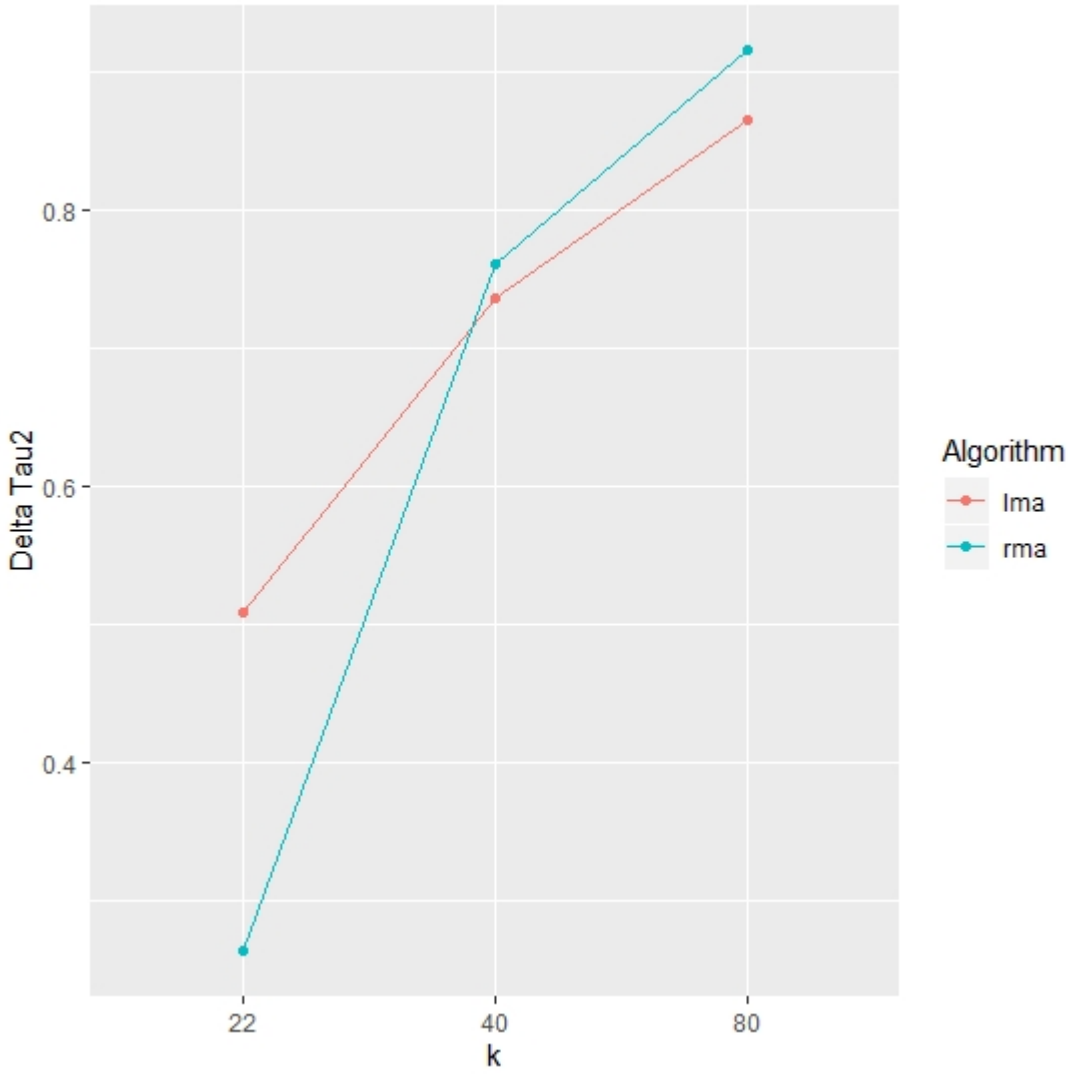
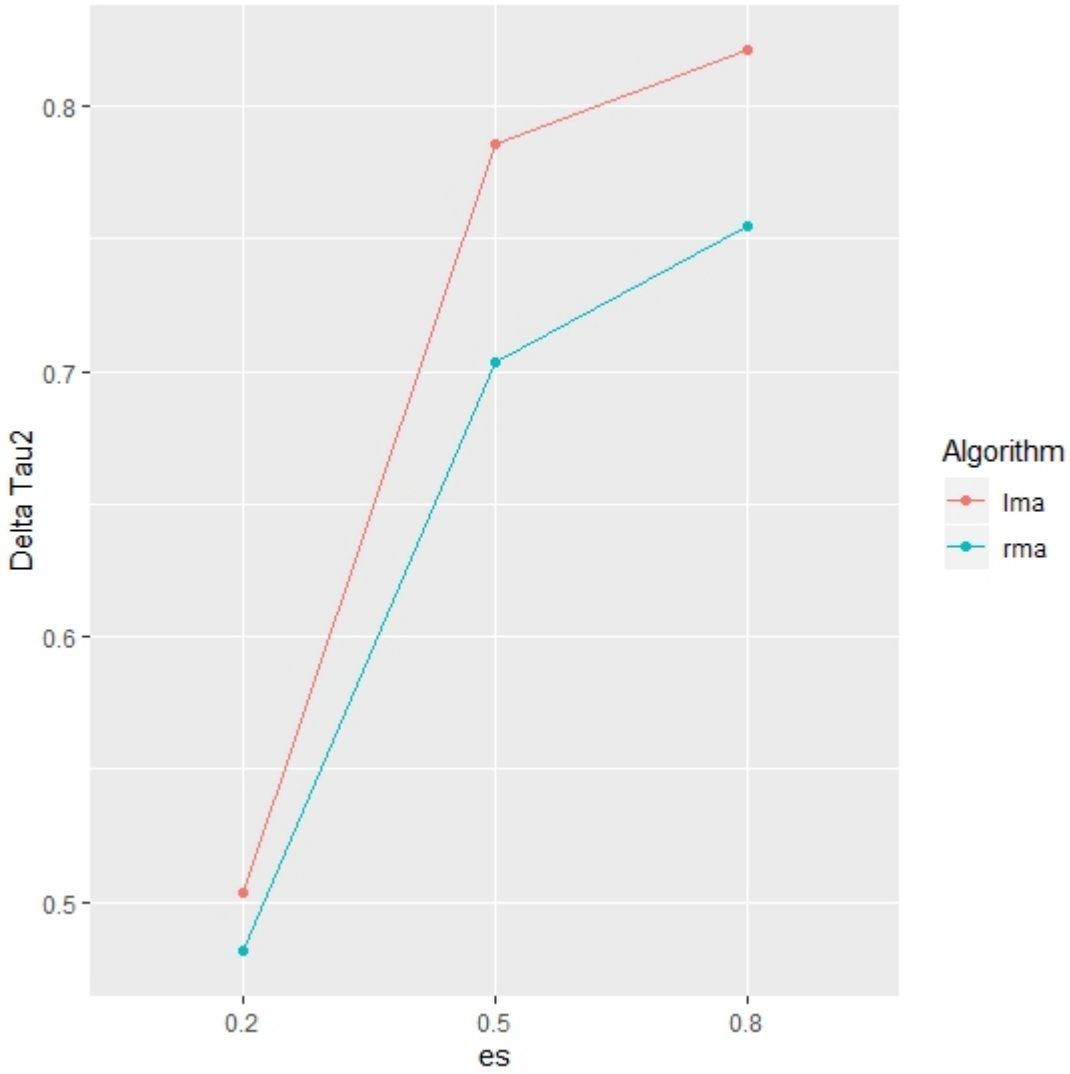


Figure 9: Interaction plot of the effect size and the algorithms on *FPS*



References

Baker, W. L., Michael White, C., Cappelleri, J. C., Kluger, J., Coleman, C. I., & From the Health Outcomes, Policy, and Economics (HOPE) Collaborative Group. (2009). Understanding heterogeneity in meta-analysis: the role of meta-regression. *International journal of clinical practice*, *63*(10), 1426-1434.

Bambra, C. (2011). Real world reviews: a beginner's guide to undertaking systematic reviews of public health policy interventions. *Journal of Epidemiology & Community Health*, *65*(1), 14 19.

Briscoe, E., & Feldman, J. (2011). Conceptual complexity and the bias/variance tradeoff. *Cognition*, *118*(1), 2-16.

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621.

Guolo, A., & Varin, C. (2017). Random-effects meta-analysis: the number of studies matters. *Statistical methods in medical research*, *26*(3), 1500-1518.

Hardy, R. J., & Thompson, S. G. (1996). A likelihood approach to meta-analysis with random effects. *Statistics in medicine*, *15*(6), 619-629.

Hedges, L. V., & Vevea, J. L. (1998). Fixed-and random-effects models in meta-analysis. *Psychological methods*, *3*(4), 486.

Hesterberg, T., Choi, N. H., Meier, L., & Fraley, C. (2008). Least angle and ℓ1 penalized regression: A review. *Statistics Surveys*, *2*, 61-93.

Higgins, J. P., Thompson, S. G., & Spiegelhalter, D. J. (2009). A re-evaluation of random-effects meta-analysis. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, *172*(1), 137-159.

Higgins, J. P., & Thompson, S. G. (2002). Quantifying heterogeneity in a meta-analysis. *Statistics in medicine*, *21*(11), 1539-1558.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning* (Vol. 112, p. 18). New York: springer.

Kitchenham, B. (2004). Procedures for performing systematic reviews. *Keele, UK, Keele University*, *33*(2004), 1-26.

Van Lissa, C. J. (2017). MetaForest: Exploring heterogeneity in meta-analysis using random forests. *Open Science Framework*. https://doi.org/10.17605/OSF.IO/KHJGB

Neely, J. G., Magit, A. E., Rich, J. T., Voelker, C. C., Wang, E. W., Paniello, R. C., ... & Bradley, J. P. (2010). A practical guide to understanding systematic reviews and meta analyses. *Otolaryngology--Head and Neck Surgery*, *142*(1), 6-14.

Panityakul, T., Bumrungsup, C., & Knapp, G. (2013). On estimating residual heterogeneity in random-effects meta-regression: a comparative study. *Journal of Statistical Theory and Applications*, *12*(3), 253-265.

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

Stanley, T. D., & Doucouliagos, H. (2017). Neither fixed nor random: Weighted least squares meta regression. *Research synthesis methods*, *8*(1), 19-42.

Veroniki, A. A., Jackson, D., Viechtbauer, W., Bender, R., Bowden, J., Knapp, G., ... & Salanti, G. (2016). Methods to estimate the between-study variance and its uncertainty in meta-analysis. *Research synthesis methods*, *7*(1), 55-79.

Viechtbauer, W., López-López, J. A., Sánchez-Meca, J., & Marín-Martínez, F. (2015). A comparison of procedures to test for moderators in mixed-effects meta-regression models. *Psychological methods*, *20*(3), 360.

Viechtbauer W (2019). "metafor: Meta-Analysis Package for R". R package version 2.1-0, URL: http://CRAN.R-project.org/package=metafor.

Viechtbauer, W. (2010). Conducting meta-analyses in R with the metafor package. *Journal of statistical software*, *36*(3), 1-48.