Select relevant moderators in meta-regression using Bayesian penalization

Caspar J. Van Lissa[1,2], Andreas M. Brandmaier[3,4], & Ernst-August Doelle[1,2]

[1] Wilhelm-Wundt-University

[2] Konstanz Business School

Author Note

Add complete departmental affiliations for each author here. Each new line herein must be indented, like this line.

Enter author note here.

The authors made the following contributions. Caspar J. Van Lissa: Conceptualization, Writing - Original Draft Preparation, Writing - Review & Editing; Ernst-August Doelle: Writing - Review & Editing.

Correspondence concerning this article should be addressed to Caspar J. Van Lissa, Padualaan 14, 3584CH Utrecht, The Netherlands. E-mail: c.j.vanlissa@uu.nl

Abstract

One or two sentences providing a **basic introduction** to the field, comprehensible to a scientist in any discipline.

Two to three sentences of **more detailed background**, comprehensible to scientists in related disciplines.

One sentence clearly stating the **general problem** being addressed by this particular study.

One sentence summarizing the main result (with the words “**here we show**” or their equivalent).

Two or three sentences explaining what the **main result** reveals in direct comparison to what was thought to be the case previously, or how the main result adds to previous knowledge.

One or two sentences to put the results into a more **general context**.

Two or three sentences to provide a **broader perspective**, readily comprehensible to a scientist in any discipline.

Keywords: keywords

Word count: X

Select relevant moderators in meta-regression using Bayesian penalization

Skeleton lasso/pema paper 1.) What is Meta-analysis? 2.) What is meta-regression and how does it complement meta-analysis? a.) Introduce Moderators b.) Study Heterogeneity + Random Sampling Error (and their difference)

2.5.) Fixed vs. random effects a.) Shortcomings of fixed effect models 3.) Shortcomings of current meta regressions w.r.t. estimating coefficients and heterogeneity: a.) Small sample size / overfitting b.) Non-normal data 4.) Various methods to estimate heterogeneity (and coefficients) a.) The use of WLS and REML 5.) Intro to Frequentist linear methods/Bayesian methods and Random forests, along with their (dis-)advantages: a.) Rma: uses WLS for estimation b.) MetaForest (Random Effects): Uses Random Forest Algorithm c.) Lasso Pema: uses penalized Lasso d.) Horseshoe Pema: uses horseshoe priors 6.) Goal of the current study 7.) Means of attaining goal and evaluation of performance: a.) simulation study b.) algorithmic performance c.) design factors d.) Impact of design factors on algorithim performance e.) Hypotheses of algorithmic performances

colour coding: I colour coded the text as to know from which file the text is copied GREEN = derived from ‘Thesis_Metaforest’ BLUE = Thesis_lasso BLACK = internship_report RED = Inserted myself

introduction

intro to 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.

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 characteristics of a study or “moderators” that could have impact on the found effect size.

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.

However, when performing such an analysis, a lot of moderators are measured and it may be unclear which are relevant and which are not. 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 and predictive power 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.

Thus, there is a need of a regularization method to curtail overfitting. Least Absolute Shrinkage and Selection Operator [LASSO] (L1-norm regularization) can fulfill this role, since it has an advantage in terms of feature selection. The goal of this project is to implement L1-norm regularization in the weighted meta-regression, developing an new estimator for penalized meta-regression.

Fixed vs Random effects 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:

$y_i = \theta + \varepsilon_i$ (1)

Where θ is the true effect size of each individual study i and $\varepsilon_i \sim \mathcal{N}(0,\, \sigma_i)$ with σ_i being the sampling error or within-study variance, which is treated as a known factor.

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”:

$W_i = \frac{1}{\sigma_i^2}$ (2)

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 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:

$y_i = \theta_i + \varepsilon_i$ (3)

With $\varepsilon_i \sim \mathcal{N}(0,\, \sigma_i)$ but, in this case θ_i on itself is given by:

$theta_i = \mu + \zeta_i$ (4)

With μ being the mean of the distribution of the true effect sizes and $\zeta_i \sim \mathcal{N}(0,\, \tau^2)$ with τ^2 being the variance of the population of true effect sizes. It could also be explained as the variance between the individual studies. 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:

$W_i = \frac{1}{\sigma_i^2 + \tau^2}$ (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 τ^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 τ^2.

Meta-regression 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_i the moderators, β_i the coefficients, with p being the number of moderators:

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

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

$$y_i = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \zeta_i + \varepsilon_i \ (7)$$

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). 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). 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 heterogeneity increases, 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.

Methods for estimating residual heterogeneity 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.

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 is an iterative method and 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 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).

Intro rma 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 meta-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.

Intro 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 λ and the regression coefficients β_j. 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 λ. If the λ 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 λ is zero, the shrinkage penalty has no impact at all and it will produce the OLS estimates.

Alternative to linear model: Tree Based models An alternative that can perform variable selection, are tree-based models. 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. Still, singletree based models have some limitations. 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 & 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.

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. In van Lissa (2017), that only presented estimates of τ^2 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. 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.

Algorithms and simulation + goal study 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.

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. Simulations are useful for evaluation of new methods like MetaForest and for the comparison with alternative methods like metaCART and the classic approaches.

Performance Criteria 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. 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:

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

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

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.

Design factors 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 follow $\sim \mathcal{B}\rceil\nabla\backslash(0.5)$, 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 mean is computed by the assessing the values of the coefficients β_j, 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 σ_i^2 is formed by varying the sizes of the samples of each study. The sample sizes $n_i \sim \mathcal{N}(\bar{n}, \frac{\bar{n}}{3}$ (Van Lissa, 2017).

Data were simulated using the random-effects model, based on four models: (A) Main effect of one moderator, $\mu_i = \beta_1 x_{1i}$ (B) Two-way interaction,

$\mu_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{1i} x_{2i}$ (E) Non-linear, cubic relationship, $\mu_i = \beta_1 x_{1i} + \beta_2 x_{1i}^2 + \beta_3 x_{1i}^3$ (F) Exponential relationship, $\mu_i = \beta_1 e^{x_{1i}}$

Impact of design factors and hypotheses 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

Results

There were 20 cases out of 388800 that had missing values on all metrics for the RMA algorithm. Closer inspection showed that both the cubic and exponential model each contributed 10 times to the missing values and only when 2, 3 or 6 moderators were taken up in the model. However, since this makes up 0.005% of the data, the missing values were chosen to be omitted from further analysis. Another thing is that the two-way interaction model only has results when there were either 3,4 or 7 moderators taken up, while the other models only have results when there were 2,3 or 6 moderators. The effects of the moderators is therefore difficult to compare between the two-way interaction and the other models.

Predictive performance Because the densities for the R^2 and MSE values were skewed, it was chosen to use the median as the metric for predictive performance, rather than the mean. The spread of the data was described using the Mean Absolute Deviation [MAD], rather than the standard deviation. It was found that the Horseshoe, Lasso and RMA algorithm performed similarly overall, $R^2_{hs} = 0.51 \pm 0.36$, $MSE_{hs} = 0.21 \pm 0.18$; $R^2_{Lasso} = 0.50 \pm 0.37$, $MSE_{Lasso} = 0.21 \pm 0.19$; $R^2_{RMA} = 0.50 \pm 0.37$, $MSE_{RMA} = 0.22 \pm 0.23$. The random forest algorithm performed worst on R^2 $R^2_{rf} = 0.35 \pm 0.38$, $MSE_{rf} = 0.22 \pm 0.19$

To determine the effect of the design factors on predictive performance R^2 of all algorithms, four separate ANOVA's were performed. The effect size η^2 per condition per algorithm, including interactions can be found in table 1. Do note that the ANOVA's were performed with the normality assumption violated. The estimates serve as a guidance, rather than an non-contestable result.

Not too surprisingly, It was found that the true effect size β had the largest effect on the predictive R^2 for all algorithms. As β increased, the performance of all algorithms increased as well. However, β did seem to interact with the model that was estimated, being either a linear, two-way interaction, cubic or an exponential model. The image of the interaction is shown in image 1. For the exponential and cubic model, the increase of R^2 slows for every higher value for β. The two-way interaction model only slows the increase when the es goes from 0.5 to 0.8, while the steepest increase for the linear model is when β increases from 0.2 to 0.4. Also noteworthy is that the cubic model seems to stagnate as the effect size goes up to 0.4 and onwards, while the other models do keep increasing. There seemed to be little difference in performance between Horsehoe, Lasso and RMA, while MetaForest performed worst.

The second largest marginal effect was that of the estimated model. It seemed that all algorithms performed best under the cubic model, followed by a similar performance on the two-way interaction and exponential models. All algorithms seemed to perform worst for the linear model. There again is little difference in performance between the Pema algorithms and RMA, altough MetaForest performs worst. Image 2 shows the relationship.

There also was a moderate interaction effect between the estimated model and the amount of skewness of the data [α], especially for the Pema algorithms. Again, Pema and RMA algorithms performed best, followed by MetaForest in all conditions. Most obvious to note is that all, except the linear model seem to benefit when the R^2 was estimated on more skewed data, although the algorithms do perform worse for the interaction model and α goes from 5 to 10. The cubic model performs best overall, but is getting caught up by the exponential and interactoin model as α increases. Image 3 shows the relationship.

The true residual heterogeneity τ^2 had a negative linear relationship for all algorithms on R^2. i.e. as τ^2 increases, R^2 decreased. Image 4 shows the relationship.

The mean sample size per study [$\bar{n}$] also had a moderate effect. For all algorithms the effect of $\bar{n}$ was positively linearly related to R^2 with the Pema and RMA algorithms performing best and Metaforest performing worst. Image 5 shows the relationship.

An especially large effect was found for the number of studies used in the training data for MetaForest, $\eta^2 = 0.27$, while this effect was substantially smaller for RMA, $\eta^2 = 0.11$; Lasso, $\eta^2 = 0.06$ and; Horseshoe, $\eta^2 = 0.05$. The relationship is positively linear for all algorithms, but the slope is especially steep for MetaForest. Image 6 shows the relationship.

Finally, the number of moderators did not have a big effect for the Pema algorithms, $\eta^2 = 0.01$ for both Horseshoe and Lasso while for RMA and Metaforest $\eta^2 = 0.05$. This relationship is generally negative with more moderators meaning worse performance, although an increase can be observed as the number of moderators increase from 4 to 6 for all algorithms except Metaforest. This increase in performance for Metaforest appears when the number of moderators go from 3 to 4. Image 7 and η^2 values indicate that the Pema algorithms are more robust against number of moderators than RMA and MetaForest.

Estimating residual heterogeneity (τ^2) The ability of the algorithms to correctly estimate τ^2 was operationalized by subtracting the true value for τ^2 from the τ^2 estimated by the algorithms. Again, the median and MAD were used as metrics for performance. The RMA algorithm showed the best results, $\Delta\tau^2_{RMA} = 0.02 \pm 0.06$, followed by the Metaforest algorithm $\Delta\tau^2_{MF} = 0.09 \pm 0.13$. The Pema algorithms performed worst $\Delta\tau^2_{Lasso} = 0.23 \pm 0.18$; $\Delta\tau^2_{hs} = 0.23 \pm 0.17$. The finding that all medians are positive implies that all algorithms have a bigger tendency to overestimate τ^2 than underestimate it. One comment to make is that uncertainty of the estimates generally increased as $\Delta\tau^2$ also increased. This implies that there is more variation in performance as median performance worsens.

To determine the effect of the design factors on estimated τ^2 of all algorithms, four separate ANOVA's were performed. The effect size η^2 per condition per algorithm, including interactions can be found in table 2. Again, the assumption of normality was violated.

The estimated model had one of the biggest effect on the correct estimation of τ^2. This is mainly because the algorithms overestimate τ^2 quite severely when the model contains cubic relationships. Image 8 shows the marginal relationship of the estimated model on $\Delta\tau^2$. It becomes more clear why this overestimation occurs when showing the interaction between the effect size and the model estimated on $\Delta\tau^2$ shown in image 9. First note the general trend which seems to be that for all models, except the linear, τ^2 gets more overestimated the higher the true effect size is. This is not the case for the linear algorithm, where the effect of β on $\Delta\tau^2$ seems close to zero, except for MetaForest. However, note the scales for the y-axes. While the interaction, linear and stay well within a confined interval, the algorithms severely overestimate τ^2 when the model contains cubic terms. Especially Metaforests overestimation is substantial when $\beta = 0.8$ and the estimated model is cubic with $\Delta\tau^2_{MF} = 2.92 \pm 2.37$, although all other algorithms also have a $\Delta\tau^2 > 1$ in these conditions. Interestingly, the Pema algorithms even outperform the RMA algorithm within these conditions. The marginal effects of β on $\Delta\tau^2$ are shown in image 10. MetaForest seems most affected by the increase in β, but in general performs better than the Pema algorithms when $\beta < 0.8$. The RMA algorithm performs best overall.

The marginal effect of the skewness of the data α on $\Delta\tau^2$ seems rather minimal. There is a slight decrease in $\Delta\tau^2$ as α increases, but it is hardly noticeable. However, this decrease becomes more clear when we add the interaction with the estimated model. Image 12 shows the conditional relationship. The algorithms seem rather unaffected by α for the linear model, and a small general decrease in $\Delta\tau^2$ as α increases can be seen in the exponential model. When the interaction model is estimated however, the algorithms seem to benefit as α increases, while for the cubic model, $\Delta\tau^2$ first increases as α increases, and afterwards decreases. RMA performs best, followed by Metaforest. The Pema algorithms perform similar, but worst.

The effect of the true τ^2 on $\Delta\tau^2$ also seemed rather unnoticeable for the RMA and Metaforest algorithm. The tendency for the Pema algorithms on the other hand, was to overestimate τ^2 more as the true τ^2 increased. Image 12 shows the marginal relationship.

The effect of the number of moderators on $\Delta\tau^2$ was not that large either. A small increase in $\Delta\tau^2$ can be seen in the RMA and Metaforest algorithm as the number of moderators increase, which is not found for the Pema algorithms. However, A small note that should be made is that the Metaforest algorithm does substantially increase in $\Delta\tau^2$ as more moderators are added when the estimated model is cubic. Image 13 shows the interaction between the number of moderators and the estimated model on $\Delta\tau^2$.

The number of studies used in the training data κ only had a substantial effect for the Metaforest algorithm. For Metaforest, the $\Delta\tau^2$ decreased quite rapidly if κ increased, especially when the cubic model was estimated. For the other algorithms, decreasing κ seemed to have little to no effect on correctly estimating the residual heterogeneity. Image 14 shows the relationship.

Finally, the average number of observations in the studies did not have a substantial effect on $\Delta\tau^2$ Image 15 shows the marginal relationship.

Variable selection To determine the extent to which the algorithms could perform correct variable selection, the proportion true positives [TP] and true negatives [TN] were calculated. There were no differences in variables selected by Highest Density Intervals or Confidence Intervals for both Lasso and Horshoe and so for both algorithm it did not matter whether the Highest Density Interval or Confidence Interval were analyzed. The mean proportions along with the standard deviations on TP and TN are provided. It was found that Metaforest had the highest proportion true positives: $TP_{mf} = 0.98 \pm 0.15$, closely followed by RMA: $TP_{RMA} = 0.96 \pm 0.21$. Horshoe performed slightly better than Lasso; $TP_{hs} = 0.91 \pm 0.29$; $TP_{Lasso} = 0.89 \pm 0.31$. As for TN, it was found that the pema algorithms performed best: $TN_{hs} = 0.93 \pm 0.16$, $TN_{Lasso} = 0.93 \pm 0.17$, followed by RMA: $TN_{RMA} = 0.89 \pm 0.20$. MF performed worst by a large margin: $TN_{mf} = 0.50 \pm 0.35$ Perfomance on TP and TN were very high for all algorithms, with all mean proportions exceeding .89. One exception was the performance on TN for MetaForest, which mean was

0.50. This implies that MetaForest had issues excluding irrelevant moderators from the models. Plots were inspected to determine the effects of the design factors on the proportions. Something different to note is that there seem to be little marginal effects of the design factors on TN, while TP is more affected. It implies that the algorithms are more robust in excluding irrelevant moderators than in including relevant moderators.

Firstly, κ only had a marginal positive effect on TP. As κ increased, TP also increased. This implies that the ability for all algorithms to attribute positive relevance to relevant moderators increased as the number of studies in the training dataset increased. MetaForest had the highest TP, followed by RMA and lastly Horshoe and Lasso. The increase in TP for higher values of κ was steeper for the pema algorithms, however. There also seemed to exist an interaction of κ with the estimated model. For the linear and interaction model, the relationship of κ seemed relatively linear. At $\kappa = 20$ the TP was relatively low for the algorithms, compared to the cubic model where TP starts at .98 and converges to 1 as κ increased. This latter relationship was also found for the exponential model, although the TP at $\kappa = 20$ was slighlty lower. Image 16 shows the interaction.

The mean observations per study had a roughly positive linear relationship with TP for all algorithms. MetaForest performed best, followed by RMA, while Horseshoe and Lasso performed worst. Image 17 shows the relationship.

The true effect size β had an interaction effect with the model estimated on TN. Only for the interaction model, TN decreased as β increased, For the other models, TN remained stable. This could be because the interaction model was only fitted when there were 3,4 or 7 moderators, while for the other models, only 2,3 or 6 moderators were used. Image 18 shows the relationship. The effect of β on TP was positive; as β increased, TP increased too.

The true τ^2 had a negative effect on TP, while skewness parameter α seemed to have little effect of both TP and TN. Image 19 shows the marginal relationship of τ^2 on TP.

Finally, there seemed to be an effect of number of moderators on TN only for the interaction model. The TN increased as the number of moderators did. Image 20 shows the relationship. The relationship was reversed for TP and was found for all algorithms, i.e. TP decreased as the number of moderators increased.

References