Author response to reviews of

Selecting relevant moderators with Bayesian regularized meta-regression

masked
submitted to *Research Synthesis Methods*

[RC] Reviewer comment Manuscript text

Dear Dr. Pigott,

Thank you for considering our manuscript for publication in *Research Synthesis Methods*. We appreciate the thorough and critical comments you solicited, which helped strengthen the argumentation, clarify the structure, and add nuance. We have attempted to address all comments as thoroughly as possible. All changes are explained below. We hope that you will find the quality of the work to be sufficiently improved.

Yours sincerely,

the authors

Associate Editor

RC: **The authors introduce their implementation of a regularisation approach for estimation and variable selection in meta-regression, which seems a timely contribution given the danger of overfitting. Some more detail or elaboration would be helpful in some places.**

AR: We have rewritten all sections of the paper and restructured some of them. Furthermore, in addressing all reviewer comments (see below), we have added detail and elaboration in many places. We hope that this addresses this comment.

RC: **please consider whether the title is appropriate, or whether (something like) "Selecting..." or "Selection of..." might be better.**

AR: We appreciate the suggestion and have changed the title to:

Selecting relevant moderators with Bayesian regularized meta-regression

RC: **in abstract and Introduction, please consider being more concrete here: it might be good to mention here already that this is about penalisation/LASSO/horseshoe priors in order to give the reader a better idea.**

AR: To address this comment and others, we have restructured the Introduction to be more to the point. We still provide the same background information in a more abbreviated form at a later point. With regard to the specific request to mention penalisation/LASSO/horseshoe priors in the Abstract and Introduction, we have made the following changes. The Abstract now reads:

> we introduce Bayesian Regularized Meta-Analysis (BRMA), which selects relevant moderators from a larger set of candidates by shrinking small regression coefficients towards zero with regularizing (LASSO or horseshoe) priors.

The Introduction now states:

> The present paper introduces *Bayesian Regularized Meta-Regression* (BRMA), an algorithm that uses Bayesian regularizing priors to perform variable selection in meta-analysis. Regularizing priors assign a high probability density to near-zero values, which shrinks small regression coefficients towards zero, thus resulting in a sparse solution. This manuscript discusses two shrinkage priors, the LASSO and regularizing horseshoe prior.

RC: on p.16 (Section "design factors"): what were the total numbers of variables (or "noise moderators")?

AR: We see now that we did not report clearly enough how many design factors there were and what they were. To address this comment and avoid confusion, we have rewritten this paragraph as follows:

> To examine performance in a range of realistic meta-analysis scenarios, seven design factors were manipulated: First, we manipulated the number of studies in the training data $k \in (20, 40, 100)$. Second, the average within-study sample size $\bar{n} \in (40, 80, 160)$. Third, true effect sizes were simulated according to two models: one with a linear effect of one moderator, $T_i = \beta x_{1i} + \epsilon_i$, and one with a non-linear (cubic) effect of one moderator, $T_i = \beta x_{1i} + \beta x_{1i}^2 + \beta x_{1i}^3 + \epsilon_i$, where $\epsilon_i \sim N(0, \tau^2)$. As both BRMA and RMA assume linear effects, simulating data from a non-linear model allows us to examine how robust the different methods are to violations of this assumption. The fourth design factor was the population effect size β in the aforementioned models, with $\beta \in (0, .2, .5, .8)$. Fifth, we manipulated the residual heterogeneity τ^2 in the aforementioned models, with $\tau^2 \in (.01, .04, .10)$. According to a review of 705 published psychological meta-analyses (Van Erp et al., 2017), these values of τ^2 fall within the range observed in practice. Sixth, we varied the number of moderators not associated with the effect size $M \in (1, 2, 5)$. These are the moderators that ought to be shrunk to zero by BRMA. Note that the total number of moderators is $M + 1$, as one moderator is used to compute the true effect size (see the third design factor). Finally, moderator variables were simulated as skewed normal moderators, with scale parameter $\omega \in (0, 2, 10)$, where $\omega = 0$ corresponds to the standard normal distribution.

RC: regarding the 3rd comment by Reviewer 1: in the context of variable selection, the so-called "penalized complexity priors" (here: exponential priors for the heterogeneity standard deviation) might play a role.

AR: To address the 3d comment by Reviewer 2, we have included the following paragraph in the revision:

The choice of prior distributions is an important decision in any Bayesian analysis. This also applies to the heterogeneity parameters. In the case of random effects meta-regression, the only heterogeneity parameter is the between-studies variance, τ^2. In the case of three-level multilevel meta-regression, there is a within-study and between-studies variance. A crucial challenge with heterogeneity parameters in meta-regression is that the number of observations at the within- and between-study level is often small. This can result in poor model convergence (Röver et al., 2021), or boundary estimates at zero (Chung, Rabe-Hesketh, & Choi, 2013). A well-known advantage of Bayesian meta-analysis is that it can overcome these challenges by using weakly informative priors, which guide the estimator towards plausible values for the heterogeneity parameters. There is less consensus, however, about which priors are most suitable for this purpose (Röver et al., 2021). BRMA uses a prior specifically developed for multilevel heterogeneity parameters (Gelman, 2006): a half-Student's t distribution with large variance, student-$t^+(3, 0, 2.5)$. Note that other relevant weakly informative priors have been discussed in the literature, such as the Wishart prior (Chung, Gelman, Rabe-Hesketh, Liu, & Dorie, 2015). There has also been increasing interest in the use of informative priors for heterogeneity parameters, which incorporate substantive knowledge about plausible parameter values (C. G. Thompson & Becker, 2020). Informative priors exert substantial influence on the parameter estimates. They thus differ from weakly informative priors, which restrict the estimator towards possible values (e.g., by excluding negative values for the variance), or guide it towards plausible values to aid model convergence. BRMA takes a pragmatic approach to Bayesian analysis, using weakly informative priors to aid convergence for heterogeneity parameters, and regularizing priors to perform variable selection for regression coefficients. The use of informative priors is out of scope for BRMA. If researchers do wish to construct alternative prior specifications, they may want to develop a custom model in `rstan` instead (Stan Development Team, 2022).

1. Reviewer 1

RC: The manuscript provides an interesting approach based on regularization to selection of covariates in meta-regression, when the number of covariates is relevant. The Authors focus on a Bayesian strategy. Although I found the topic interesting, the study needs some additional investigations and the text needs to be substantially modified in order to provide a clear and exhaustive picture of the proposal. My comments and questions are listed below.

AR: We thank the Reviewer for their constructive comments, which substantially helped improve the quality of the manuscript. We have tried to address them all in turn, as detailed below.

RC: The introduction is too long in terms of explanation of the meta-regression and selection of covariates problems, with many repetitions. For example, end of page 4 and beginning of page 5 include many repetitions about between-study heterogeneity. Other parts, instead, require a substantially deeper explanation.

AR: To address this comment and others, we have restructured the Introduction to be more to the point. We still provide the same background information in a more abbreviated form at a later point. We have elaborated on all the points that were raised by the Reviewers (see our responses to the other comments).

RC: Which theories useful for variable selection do you refer to in page 5 ? The description is extremely vague.

AR: We have attempted to clarify this section by invoking an additional reference and rewriting it as follows:

> Prior authors have stressed the need to perform variable selection in meta-regression, for example, by limiting the number of moderators considered (S. G. Thompson & Higgins, 2002). This does not resolve the problem, however, as failing to consider a moderator does not mean that it is irrelevant. Instead, moderators ought to be selected based on their theoretical or empirical relevance for the studied effect.
>
> One approach to is to select variables on theoretical grounds. An important caveat is that theories that describe phenomena at the level of individual units of analysis do not necessarily generalize to the study level. Using such theories for variable selection amounts to committing the ecological fallacy: generalizing inferences across levels of analysis (Jargowsky, 2004). The implications of the ecological fallacy for interpreting the *results* of meta-regression are well-known (Baker et al., 2009; S. G. Thompson & Higgins, 2002): For example, meta-regression may find a significant positive effect of average sample age on the effect size of a randomized controlled trial, even if age is uncorrelated with treatment efficacy within each study. Less well-known is that the same problem applies when using individual level theory to select study level moderators: If theory states that an individual's age influences their susceptibility to treatment, that does not imply that average sample age will be a relevant moderator of study-level treatment effect in meta-regression. One rare example of study level theory is the *decline effect*: effect sizes in any given tranche of the literature tend to diminish over time (Schooler, 2011). When, by coincidence, a large effect is found, it initially draws attention from the research community. Subsequent replications then find smaller effects due to regression to the mean. Based on the decline effect, we might hypothesize "year of publication" to be a relevant moderator of study effect sizes. At present, few such study level theories about the drivers of heterogeneity exist, and until they are developed, theory has limited utility for variable selection.

RC: The topic of the study is variable selection from a Bayesian point of view: but the description is so short. Just a few lines about the novelty of the paper in page 6, with no clear description of the proposal and with the reference to pema package without even the explanation of the name of the package.

AR: To address this comment, we have rewritten the opening paragraph to clearly state the problem addressed in this paper, and the proposed solution:

> A common application of meta-analysis is to summarize existing bodies of literature. A crucial challenge is that there is often substantial heterogeneity between studies, because similar research questions are studied in different labs, sampling from different populations, and using different study designs, instruments, and methods. Any of those between-studies differences can introduce *systematic heterogeneity* in observed effect sizes. Suspected causes of systematic heterogeneity can either be used as exclusion criteria, or controlled for using *meta-regression* (see López-López, Marín-Martínez, Sánchez-Meca, Van den Noortgate, & Viechtbauer, 2014). The latter approach provides an opportunity to learn which factors impact the effect size found. However, a limitation of meta-regression is that it requires a relatively high number of cases (studies) per parameter to obtain sufficient statistical power. In applied meta-analyses, the number of available studies is often low (Riley, Higgins, & Deeks, 2011). This introduces a risk of overfitting, which results in uninterpretable model parameters (Hastie, Tibshirani, & Friedman, 2009). In extreme cases, the ratio of cases to parameters can be so low that the model is not (empirically) identified, resulting in non-convergence (Akaike, 1974). Accounting for between-studies heterogeneity thus poses a non-trivial challenge to classic meta-analytic methods. The risk of arriving at false-positive conclusions when there are many potential moderators is so ubiquitous that it was referred to as the "primary pitfall" in meta-regression (S. G. Thompson & Higgins, 2002). The present study introduces a novel method to overcome this pitfall by imposing Bayesian regularizing (LASSO and regularizing horseshoe) priors on the regression coefficients. These priors shrink the effect of irrelevant predictors towards zero while leaving important predictors relatively unaffected. The result is a sparse model with fewer non-zero parameters, which benefits model convergence, reduces overfitting, and helps identify relevant between-study differences.

RC: Statistical underpinnings: Again, I found a lot of repetitions in the text.

AR: We have merged several sections to avoid repetition in the text.

RC: Probably, in (1) you mean θ and not Θ? As you have θ int the text. Why do you use theta and not beta, as you do for the meta-regression model in (7)? The notation would be simplified and would be more consistent.

AR: We have made the suggested change and now use the intercept β_0 in place of Θ. We have also simplified the notation of the estimated population effect to $\hat{\beta}_0$.

RC: The sentence in lines 132-133 is not clear.

AR: We have attempted to clarify this sentence as follows:

> In contrast to the fixed effect model, the random effects model assumes that all studies provide some information about the underlying distribution of true effect sizes. Fixed effect weights would discount the information smaller studies provide about the scale of this distribution, which is represented by it's variance τ^2. To overcome this limitation, the weights are attenuated in proportion to the variance. The random effects weights are thus given by $w_i = \frac{1}{v_i + \hat{\tau}^2}$.

RC: line 147: probably the explanation of the error term should be anticipated where the error term appears for the first time.

AR: To address this comment, we have moved the explanation of the "error term" to the paragraph where it is first discussed.

RC: line 152: "to estimate this model " has a preferable statistical sound than "to solve this model"

AR: We have made the suggested change.

RC: lines 163- 166 describe some features of regularization. . . .but regularization is introduced later, from line 167.

AR: To address this comment, we have moved this paragraph about regularization into the section on regularized regression.

RC: Regularized regression + Bayesian estimation I found both the sections very short and lacking of many details.

AR: Our aim is to provide a high-level conceptual introduction of the method's underpinnings, not to replicate the many excellent technical publications that already exist on these methods. To address this comment, we have edited all technical sections for greater clarity and detail. Additionally, we explicitly direct the reader to appropriate introductory texts that go more in depth:

> To understand how regularized regression introduces bias, consider a comparison between ordinary least squares regression and LASSO regression (for a more elaborate introduction, see Tibshirani, 2011).

And:

> For a more extensive introduction to Bayesian estimation, see McElreath (2020).

RC: what about the existing literature using lasso for meta-regression?

AR: We were not aware of any existing literature at the time of writing; most likely because several such papers were only published after we wrote our draft - but we thank the Reviewer for encouraging us to search the literature again. Based on our findings, we now include the following paragraph:

> Some seminal studies have applied the LASSO to perform moderator selection in meta-regression (Requia et al., 2018; Rosettie et al., 2021; Sebri & Dachraoui, 2021). This suggests that others have recognized its potential for exploring heterogeneity when the number of moderators is relatively high to the number of studies. However, these existing publications have taken a two-step approach, whereby moderators are first selected using LASSO regression, and selected moderators are then included in meta-regression analysis. This approach is fraught; firstly, because of known problems of inference after variable selection (Zhang, 1992). As the moderators included in the second step are based on an exploratory first step, their parameters are not valid for inference. Secondly, although the LASSO model in the first step accounts for potential multicollinearity by including all colinear variables but restricting the size of their coefficients, the meta-regression in the second step no longer does so. A three-step extension of the two-step approach exists that uses principles from the causal inference literature to overcome these limitations (Belloni, Chernozhukov, & Hansen, 2014). BRMA, by contrast, overcomes these limitations by introducing a one-step approach that performs inference within the penalized framework.

RC: In addition, I expected some references about the theory of LASSO given the relevance of the instrument in the literature.

AR: To address this comment, we now cite the following theoretical paper:

> To understand how regularized regression introduces bias, consider a comparison between ordinary least squares regression and LASSO regression (for a more elaborate introduction, see Tibshirani, 2011).

RC: line 183-184: "other penalties exist": ok, so discuss about, indicate advantages and disadvantages of the other solutions; why do you not focus on them in the study? If you do not, just explain why.

AR: We respectfully feel that indicating advantages and disadvantages of other penalties is out of scope for the present paper. The only reason we explain the LASSO penalty is to explain how regularization works. Since we ultimately do not use a frequentist solution with shrinkage penalties, there is little point to explaining other penalties in detail.

To address this comment, we now clarify why we explain the LASSO penalty:

> Note that many other regularizing penalties exist. This introduction focuses on the LASSO penalty because it is most ubiquitous, easy to understand, and has an analogue in Bayesian estimation, as explained in the next section.

RC: Bayesian estimation: I think that some additional methodological details and additional explanations about priors and their comparison to the classical LASSO would help the reader.

AR: As requested, we have added this information. These sections now read:

> An alternative to the use of a shrinkage penalty is Bayesian estimation with a regularizing prior. Whereas the aforementioned (frequentist) approaches treat every possible parameter value as equally plausible, Bayesian estimation combines information from the data with a *prior distribution* that assigns a-priori probability to different parameter values. Likely parameter values have a high probability density, and unlikely parameter values have a low probability density. The prior distribution is updated with the likelihood of the data to form a posterior distribution, which reflects expectations about likely parameter values after having seen the data. For a more extensive introduction to Bayesian estimation, see McElreath (2020).
>
> A regularizing prior distribution shrinks small coefficients towards zero by assigning high probability mass to near-zero values. There are many different regularizing prior distributions, some of which are analogous to specific frequentist methods (van Erp, Oberski, & Mulder, 2019). For example, a double exponential prior (hereafter: LASSO prior) results in posterior distributions whose modes are identical to the estimates from LASSO-penalized regression (Park & Casella, 2008). Both the the frequentist LASSO penalty and the Bayesian LASSO prior have a tuning parameter λ that controls the amount of regularization. In frequentist LASSO, its value is usually chosen via cross-validation (Hastie et al., 2009). In the Bayesian approach, by contrast, its value can be optimized during model estimation.
>
> One limitation of the LASSO prior is that it biases all regression coefficients towards zero - for relevant as well as irrelevant moderators. To overcome this limitation, regularizing priors with better shrinkage properties have been developed. These priors still pull small regression coefficients towards zero, but exert less bias on larger regression coefficients. One example is the horseshoe prior (Carvalho, Polson, & Scott, 2010). It has heavier tails than the LASSO prior, which means that it does not shrink (and therefore bias) substantial coefficients as much. One limitation of the horseshoe prior is that it is difficult to specify a prior that ensures a sufficiently sparse solution. The regularizing horseshoe was introduced to overcome this limitation (Piironen & Vehtari, 2017b).

RC: Why details relevant for the Bayesian estimation are included in the "Implementation" section? I mean lasso prior and horseshoe prior, that is, details useful to understand the methodology.

AR: We agree with the Reviewer that this information would be better positioned in another section. To address this comment, we have restructured the paper; the information about priors is now incorporated into the *Bayesian estimation* section, and the *Implementation* section now exclusively relates to the `pema` R-package.

RC: Can you comment about the choice of the numerical values in the priors at the end of page 11?

AR: We now explain this as follows:

> For both Bayesian priors, we used default values proposed in prior literature, see Table 1. Default values for the LASSO prior were based on van Erp et al. (2019), and default values for the regularizing horseshoe prior were based on Piironen and Vehtari (2017a).

Note that the results of our simulation study again validate the fact that these are sensible defaults.

RC: line 221: "this extension": which one?

AR: We recognize that this choice of words was confusing. The regularizing horseshoe prior is an extension of the horseshoe prior. To address this comment, we now explain this in the text:

> One limitation of the LASSO prior is that it biases all regression coefficients towards zero - for relevant as well as irrelevant moderators. To overcome this limitation, regularizing priors with better shrinkage properties have been developed. These priors still pull small regression coefficients towards zero, but exert less bias on larger regression coefficients. One example is the horseshoe prior (Carvalho et al., 2010). It has heavier tails than the LASSO prior, which means that it does not shrink (and therefore bias) substantial coefficients as much. One limitation of the horseshoe prior is that it is difficult to specify a prior that ensures a sufficiently sparse solution. The regularizing horseshoe was introduced to overcome this limitation (Piironen & Vehtari, 2017b).

RC: line 225: "This is accomplished by.." Why?

AR: We have rephrased this so that it is clearer that the formula explains how this is achieved. The section now reads:

> An attractive property of this shrinkage prior is that it can incorporate prior information regarding the expected number of relevant moderators. This is accomplished by calculating the scale of the global shrinkage parameter λ_0^2 based on the expected number of relevant moderators p_{rel}. The shrinkage parameter is then given by $\lambda_0^2 = \frac{p_{rel}}{p - p_{rel}} \frac{\sigma}{\sqrt{n}}$, where σ is the residual standard deviation and n equals the number of observations.

RC: line 240: "values are reasonable in most applications" Why? Can you prove this?

AR: We thank the Reviewer for pointing out that this statement raises more questions than it answers. To address this Reviewer comment, we have moved this statement to the Discussion and slightly rephrased it to clarify that it is supported by the results of the present study. It now reads:

> All simulations were conducted with default settings for the model's prior distributions, based on prior research (Piironen & Vehtari, 2017b; van Erp et al., 2019). Our results suggest that these defaults are suitable for a wide range of situations, including when model assumptions are violated. However, bear in mind that model parameters are influenced by the prior distribution. It is good practice to perform sensitivity analysis to determine how sensitive the model results and inferences are to different prior specifications. Performing sensitivity analyses is particularly important when the sample is small, as in this case, the prior is more influential.

RC: lines 244-251 seems to be pertinent to the methodology of the proposal not to implementation.

AR: We agree with the Reviewer that the writing in this section was too methodological. To address this comment, we have made two changes: This paragraph has been moved to the section on Bayesian estimation, and we have rewritten it as follows:

> The frequentist LASSO algorithm shrinks coefficients to be exactly equal to zero, and thus inherently performs variable selection. Other approaches to regularization - frequentist or Bayesian - lack this property. However, an advantage of the Bayesian approach is that its posterior distributions lend itself to exact inference. One can use probability intervals to determine which population effects are likely non-zero; for example, by selecting moderators whose 95% interval excludes zero. Two commonly used Bayesian probability intervals are the credible interval and the highest posterior density interval (McElreath, 2020). The credible interval (CI) is the Bayesian counterpart of a confidence interval, and it is obtained by taking the 2.5% and 97.5% quantiles of the posterior distribution. The highest posterior density interval (HDPI) is the narrowest possible interval that contains 95% of the probability mass (Kruschke, 2015). When the posterior distribution is symmetrical, the CI and HDPI are the same. However, when the posterior is skewed, the HPDI has the advantage that all parameter values within the interval have a higher posterior probability density than values outside the interval. This suggests that the HPDI might be superior for performing inference on residual heterogeneity parameters, which have a skewed posterior distribution by definition. For inference on regression coefficients, the choice of interval is likely less crucial.

RC: More in general, I do not like to see a mixture of methodology and R code. I think I would be more useful for the reader to have sections describing the proposed approach and the underlying methodology (something that is not present at the emano) and then having an appendix about the code.

AR: We see that a stricter separation of methodology and code could aid readability. Therefore, we have addressed this comment by relegating all code to the Implementation section.

We do believe that there are advantages to having code in the paper and not in an appendix, as the intended readership for this paper consists of those who wish to conduct a Bayesian regularized meta-analysis. For this purpose, it is important that readers understand how to translate the methodological principles to a model specified in the `pema` package. We hope that the current revision strikes a reasonable compromise.

RC: line 278: "x==0" could be substituted by "x is equal to 0" as this is a text, not a source code.

AR: We have made the suggested change.

RC: Intercept. I don't think it is so relevant to deserve a section. Why not inserting a line in the section devoted to the code?

AR: This section is important because omitting the intercept from the model has very different implications for regularized meta-regression than for non-regularized models.

To clarify this, we have rewritten this section to focus on the implications of including or excluding the intercept. The line explaining *how* to include or exclude the intercept has been moved to the Implementation paragraph, consistent with the Reviewer's request to relegate code to a separate section:

The general linear model used in BRMA typically includes an intercept, which reflects the expected value of the outcome when all predictors are equal to zero, and regression coefficients for the effect of moderators. If the analysis contains categorical predictors, it may be desirable to omit this intercept. To understand why, first consider the model with an intercept. Standard practice is to encode category membership with dummy variables, with values $x \in \{0, 1\}$. For a variable with c categories, the number of dummy variables is equal to $c - 1$. The omitted category functions as a reference category, and its expected value is represented by the model intercept b_0. The regression coefficients of the dummy variables, $b_{1...c}$, indicate the difference between the expected values of the reference category and of the category represented by the dummy. This is useful when there is a meaningful reference category. For example, imagine a study on the effectiveness of interventions for specific phobia with two interventions: Treatment as usual, and a novel intervention. In this case, it makes sense to code treatment as usual as the reference category, and dummy-code the new contender. The intercept b_0 then represents the average effect size of treatment as usual, and the effect of the dummy b_1 indicates whether the newly developed intervention has a significantly different effect size from treatment as usual. In other cases, there may not be a straightforward reference category. For example, imagine a study on the effectiveness of one intervention for specific phobia in two continents. In this case, it makes more sense to estimate the average effect for all continents separately - in other words, to conduct a multi-group analysis. This is achieved by removing the intercept, and including all c dummy variables. In the context of standard linear regression, both approaches are equivalent, but in regularized regression, shrinkage affects the intercept differently from the dummy variables. Consequently, a reasoned choice must be made about whether to include an intercept or not.

RC: Performance indicators, Design factors, Predictive performance * You split the data in training set and test set, for each iteration. So, how do you account for the variability associated to the choice of the test set (as done for example in cross-validation)?

AR: We respectfully believe that this comment is based on a misunderstanding of the procedure. The Reviewer's comment seems to suggest that we split a finite sample into a training and testing set. This is not the case.

Instead, we simulate both data sets from the same known data generating model. Each sample from this model is thus independent and identically distributed (IID). The distinction between training and testing is arbitrary. This is different from real data analyses, where the available dataset is finite, and training and testing data are drawn from that finite dataset - thus violating the assumption of IID.

In a real dataset, if one case has very unusual values (i.e., an outlier), the model could be biased if this case ends up in the training data, and the performance metrics could be biased if this case ends up in the testing data. In our simulation study, outliers have exactly equal probabilities of appearing in the training or testing data - and as we are not sampling from a finite dataset, the probability of having an outlier in the training data is independent from having one in the testing data.

Their effect on the estimated parameters is known as "Monte Carlo error" and averages out across replications of the simulation.

RC: I'm surprised by the large number fo linea devoted to the explanation of such a basic indicator like R2. In addition, it is well known to be optimistic, and thus the adjusted version I usually preferable. Why don't you evaluate the method using other indicators, such as the adjusted r2, AIC, or BIC?

AR: We respectfully feel that it is important to define how we operationalized each outcome metric, and why we use it. In fact, it seems that our explanation was not sufficiently clear, because the outcome metric is not

the basic indicator R2, as the Reviewer has apparently understood, but rather, a different type of R2 that is suitable for assessing overfitting.

To address this comment, we have tried to explain the outcome metric, R^2_{test} more clearly:

> Our primary performance indicator was predictive performance, a measure of model generalizability. To compute it, for each iteration of the simulation, both a training data set and a testing data set are generated from the same known population model. The number of cases in the training data vary according to the design factors of the simulation study. The number of cases in the testing data set was always 100. The models under evaluation (BRMA, RMA) were estimated on the training data, and used to predict cases in the testing data. Predictive performance was operationalized as the model's explained variance in the testing data, R^2_{test}, calculated as follows:
>
> $$R^2_{test} = 1 - \frac{\sum_{i=1}^{k}(y_{i,test} - \hat{y}_{i,test})^2}{\sum_{i=1}^{k}(y_{i,test} - \bar{y}_{train})^2}$$
>
> Where k is the number of studies in the testing data set, $\hat{y}_{i-test}$ is the predicted effect size for study i, and $\bar{y}_{train}$ is the mean of the training data. The R^2_{test} differs from the familiar R^2 metric: R^2 describes the proportion of variance a model explains in the training data, and it always increases as the model becomes more complex. By contrast, R^2_{test} reflects the explained variance in the testing data. Remember that BRMA was developed to reduce the risk of overfitting meta-regression models. The R^2_{test} is a useful metric to detect overfitting, which causes it to decrease, or even become negative.

With regard to the Reviewer's suggestion to use AIC or BIC: the reason we do not use these is because they are comparative fit indices, only suitable for comparing a set of models; we have only a single model and AIC or BIC do not speak to the absolute fit of a single model.

The reason we do not use adjusted R^2 is because it is an approximation of R^2_{test} that can be calculated on the training data. Since we can simulate testing data from a known model, we do not need to *estimate* R^2_{test} - we can *calculate* it directly.

RC: line 344: I suppose you mean the estimate of sensitivity and specificity.

AR: We have rephrased this as suggested.

RC: line 368: 100 datasets only? Probably, there is a typo, as 100 is the number of datasets included in the test set, if I'm not wrong. In any case, 100 is a very low number of replicates for a simulation study, I suggest to increase at least to the common value equal to 1,000.

AR: As reported, we used 100 replications per simulation condition because these simulations are very computationally demanding, and increasing the number of replications adds only marginal accuracy gains. In total, we thus simulated 194.400 cases. Many of the analyses marginalize over some design factors, thus we always have a very large sample size. We do not see an urgent need to increase the number of replications, but we defer to the Editor in this matter.

RC: line 385: do you mean "Table" 1?

AR: Thank you for pointing out this omission; we have corrected it.

RC: line 405: "the" most negative?

AR: We corrected this, and have rephrased the sentence for clarity.

RC: lines 449-452 are not necessary as bias and variance are known concepts.

AR: We respectfully feel that it is prudent to define how all outcome metrics are operationalized. Even though the general concepts of bias and variance may be known to most readers, we cannot assume that they will know exactly how these are calculated in the context of this simulation study - particularly for the broad readership of a journal like Research Synthesis Methods. See also our response to the comment about defining R^2_{test}

RC: However, in Tables I can see only bias, and not information about variance.

AR: This is correct; we had not reported these statistics for two reasons:

1. Unlike bias, they must be computed on aggregate across replications of each simulation condition. This results in a strange situation where each condition has only one outcome value. But it is probably fine to interpret these effect sizes as descriptive statistics.
2. The differences between algorithms in variance of τ^2 were negligible. We did not want to devote two tables to, essentially, a matrix of zeroes.

To address this comment, we now describe the results in the text, and provide the requested tables as online Supplements:

> The variance of parameter estimates cannot be calculated on a case-by-case basis. Instead, it is calculated across replications for each simulation condition. Across simulation conditions, parameters estimated via HS had the lowest variance, $V_{HS} = 0.32$, followed by LASSO, $V_{LASSO} = 0.34$, and then RMA, $V_{RMA} = 0.38$. Online Supplemental Table S1 provides an overview of the effect size of design factors on variance of the regression coefficients. Notably, the differences between algorithms are very small; the largest effect sizes were observed for the difference between HS and RMA in the effects of effect size, sample size, and model, all with $\eta^2 < 0.01$.

And:

> The variance of the residual heterogeneity was calculated across replications for each simulation condition. The LASSO estimates of τ^2 had the lowest variance, $V_{LASSO} = 1.47$, followed by HS, $V_{HS} = 1.50$, and then RMA, $V_{RMA} = 1.71$. Online Supplemental Table S2 provides an overview of the effect size of design factors on variance of the residual heterogeneity. All differences between algorithms were small, $\eta^2 \leq 0.002$.

RC: lines 469-470: there is a white space

AR: We thank the Reviewer for pointing this out; a fragment of a sentance was cut here. We restored the original text.

RC: Could you please avoid writing pema::bonapersona and use instead dataset titled bonapersona in the R pema package? This is not a list of lines code, but a text.

AR: We have done as requested.

RC: you have 440 experiments in the dataset: do you mean 440 study included in the meta-regression? If so, the number is much. Larger than the scenarios evaluated in the simulations.

AR: This observation is correct, but this is not a problem. The method is validated for up to 100 studies, so it should also be valid for >100 studies, as a larger sample size gives more power.

The reason we used these data is because we were able to secure permission from the original author to include them in the R-package. This means we can make the examples reproducible, requiring only the `pema` package. We feel that it adds a lot to the paper to present a reproducible real-data example based on a novel data set. To address this comment, we now acknowledge in the text that this sample is relatively large compared to the sample sizes of the simulation study:

> This meta-analysis of over 400 experiments investigated the effects of early life adversity on cognitive performance in rodents. Note that the sample is much larger than the maximum used to validate BRMA in our simulation study. As larger samples provide greater statistical power, it should also be valid for this sample.

RC: Finally, There are many typos and inconsistencies throughout the text (e.g., data set and dataset, random effect and randoms-effects, . . .).

AR: We have proofread and spell checked the manuscript, and corrected all mistakes found.

2. Reviewer 2

RC: Comments to the Author The authors investigated selection of moderators for the meta-regression, which is especially useful to explain the sources of heterogeneity between trials. They proposed the use of regularizing priors for meta-regression within a Bayesian framework. They shared publicly available R package "pema" which relies on rstan, which helps practitioners to use the proposed methods. I think the contribution of the paper and R package is important. However, I have some comments.

AR: We thank the Reviewer for their encouraging words and thought-provoking comments, which we feel have helped improve the quality of the manuscript.

[RC 2.1.] Major points

RC: 1) I think it is important to note some limitations of meta-regression. For example, the use of moderators might "break" the randomization, when the studies analyzed are randomized controlled trials. This is because it is not possible to randomize patients to one moderator, see Thompson and Higgins (2002) section 3 for further discussion of limitations of meta-regression of clinical trials.

AR: We agree with the Reviewer that there are important limitations to meta-regression, and Thompson & Higgins provide an excellent analysis thereof. Most of the limitations and pitfalls they discuss come down to two fundamental points which are already addressed throughout our paper: 1) the curse of dimensionality and its implications for multicolinearity (limitations i, iv, vii and pitfalls i, ii, iii, and iv), and 2) the ecological fallacy (limitation ii and iv). They further address three singular points which cannot be attributed to the two fundamental points: Limitation iii relates to the problem of restriction of range in moderators (which we do not address), Limitation vi relates to limited data availability (which we address in our discussion of missing data), and Limitation v makes a point about regression to the mean in RCTs that we fail to grasp.

Since our method is a statistical solution to point 1), we devote most attention to it. We also acknowledge most of the other points. However, we respectfully feel that it is out of scope for the present paper to devote further attention to limitations of meta-regression that our method does *not* address; especially since Thompson & Higgins already did so.

To address this comment, we have referenced the Thompson & Higgins paper in several places, and we have added the following sentences to our Recommendations for Applied Researchers:

> Before conducting meta-regression, researchers should be aware of its limitations (see S. G. Thompson & Higgins, 2002). These can be subdivided into four categories: 1) the curse of dimensionality and its corrolary implications for multicolinearity; 2) the ecological fallacy; 3) limited information on moderator variables, including missing data and restrictions of range. BRMA seeks to address the first of these limitations, because the problems that arise from meta-analyzing small and heterogeneous bodies of literature are so ubiquitous that they have been referred to as the primary pitfall in meta-regression (S. G. Thompson & Higgins, 2002). Nonetheless, all applicable limitations should be acknowledged in the resulting publication.

RC: 2) Important references are missing, in which the use of regularization or penalization approaches for the meta-analysis were investigated, although not the same purpose as in the current paper. For example, in Chung et al (2013) regularizing prior was used for the heterogeneity parameter to avoid boundary estimates, and in Günhan et al (2020) regularizing priors were used for the heterogeneity parameter and treatment effect parameter to deal with data sparsity. Finally and most importantly, Röver et al (2021) reviewed different use of regularizing priors and provided some guidance. Note that in the mentioned references, the term of weakly informative priors are used instead of regularizing priors (also see an earlier reference by Gelman (2006) in a hierarchical model context). Mention of these publications and relationship to the present paper can help the reader.

AR: First of all, we thank the Reviewer for suggesting these references, which had eluded our literature search. We now cite them in relevant places in the manuscript.

Second of all, we agree that the use of weakly informative priors for heterogeneity parameters was insufficiently discussed in the previous version of our manuscript, and have now added a paragraph on that topic (see our response to the next comment).

Third of all, we feel that it is important to stress that the use of weakly informative priors (WIP) for heterogeneity parameters serves a categorically different purpose than the use of regularizing priors for regression coefficients - even if the term "regularization" has been used for both in prior literature. The purpose of WIP for heterogeneity parameters is to aid model convergence and avoid boundary estimates. The purpose of regularizing priors for regression coefficients is to perform variable selection. We believe that the newly added paragraph sufficiently clarifies these distinct uses of "regularization" (see response to next comment).

RC: 3) An important part of Bayesian random-effects meta-analysis is the specification of the prior for the heterogeneity parameter tau. Can you include more specifications on the prior choice for tau and influence of the choice to the results (if there is any).

AR: We agree with the Reviewer that the prior specification for the heterogeneity parameter tau is important and should be addressed in the text. To address this comment, we have included the following paragraph in the revision:

The choice of prior distributions is an important decision in any Bayesian analysis. This also applies to the heterogeneity parameters. In the case of random effects meta-regression, the only heterogeneity parameter is the between-studies variance, τ^2. In the case of three-level multilevel meta-regression, there is a within-study and between-studies variance. A crucial challenge with heterogeneity parameters in meta-regression is that the number of observations at the within- and between-study level is often small. This can result in poor model convergence (Röver et al., 2021), or boundary estimates at zero (Chung et al., 2013). A well-known advantage of Bayesian meta-analysis is that it can overcome these challenges by using weakly informative priors, which guide the estimator towards plausible values for the heterogeneity parameters. There is less consensus, however, about which priors are most suitable for this purpose (Röver et al., 2021). BRMA uses a prior specifically developed for multilevel heterogeneity parameters (Gelman, 2006): a half-Student's t distribution with large variance, student-$t^+(3, 0, 2.5)$. Note that other relevant weakly informative priors have been discussed in the literature, such as the Wishart prior (Chung et al., 2015). There has also been increasing interest in the use of informative priors for heterogeneity parameters, which incorporate substantive knowledge about plausible parameter values (C. G. Thompson & Becker, 2020). Informative priors exert substantial influence on the parameter estimates. They thus differ from weakly informative priors, which restrict the estimator towards possible values (e.g., by excluding negative values for the variance), or guide it towards plausible values to aid model convergence. BRMA takes a pragmatic approach to Bayesian analysis, using weakly informative priors to aid convergence for heterogeneity parameters, and regularizing priors to perform variable selection for regression coefficients. The use of informative priors is out of scope for BRMA. If researchers do wish to construct alternative prior specifications, they may want to develop a custom model in `rstan` instead (Stan Development Team, 2022).

RC: 4) In the abstract and in the main text, it states "We present a simulation study to validate the performance of BRMA relative to state-of-the-art meta-regression (RMA)". What does RMA refer to here, is it "Random effect Meta-Analysis using REML"? If yes, I think the term random-effect meta-analysis (or meta-regression) using RMLE is more clear than state-of-the-art meta-analysis (or meta-regression).

AR: We have made the suggested correction. Both in the Abstract and in the text, the first time we introduce RMA, we introduce it as "random effects meta-analysis using restricted maximum likelihood".

[RC 2.2.] Minor points

RC: Line 385: "see 1": I think Table is missing

AR: Thank you; we have corrected this.

3. References

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

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. https://doi.org/10.1111/j.1742-1241.2009.02168.x

Belloni, A., Chernozhukov, V., & Hansen, C. (2014). Inference on Treatment Effects after Selection among High-Dimensional Controls. *The Review of Economic Studies*, *81*(2), 608–650. https://doi.org/10.1093/restud/rdt044

Carvalho, C. M., Polson, N. G., & Scott, J. G. (2010). The horseshoe estimator for sparse signals. *Biometrika*, *97*(2), 465–480. https://doi.org/10.1093/biomet/asq017

Chung, Y., Gelman, A., Rabe-Hesketh, S., Liu, J., & Dorie, V. (2015). Weakly Informative Prior for Point Estimation of Covariance Matrices in Hierarchical Models. *Journal of Educational and Behavioral Statistics*, *40*(2), 136–157. https://doi.org/10.3102/1076998615570945

Chung, Y., Rabe-Hesketh, S., & Choi, I.-H. (2013). Avoiding zero between-study variance estimates in random-effects meta-analysis. *Statistics in Medicine*, *32*(23), 4071–4089. https://doi.org/10.1002/sim.5821

Gelman, A. (2006). Prior distributions for variance parameters in hierarchical models (comment on article by Browne and Draper). *Bayesian Analysis*, *1*(3), 515–534. https://doi.org/10.1214/06-BA117A

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: Data mining, inference, and prediction* (Second). New York: Springer.

Jargowsky, P. A. (2004). The Ecological Fallacy. In K. Kempf-Leonard (Ed.), *The Encyclopedia of Social Measurement* (Vol. 1, pp. 715–722). San Diego, CA: Academic Press.

Kruschke, J. K. (2015). Chapter 12 - Bayesian Approaches to Testing a Point ("Null") Hypothesis. In J. K. Kruschke (Ed.), *Doing Bayesian Data Analysis (Second Edition)* (pp. 335–358). Boston: Academic Press. https://doi.org/10.1016/B978-0-12-405888-0.00012-X

López-López, J. A., Marín-Martínez, F., Sánchez-Meca, J., Van den Noortgate, W., & Viechtbauer, W. (2014). Estimation of the predictive power of the model in mixed-effects meta-regression: A simulation study. *British Journal of Mathematical and Statistical Psychology*, *67*(1), 30–48. https://doi.org/10.1111/bmsp.12002

McElreath, R. (2020). *Statistical Rethinking: A Bayesian Course with Examples in R and STAN* (Second). Boca Raton, FL: CRC Press.

Park, T., & Casella, G. (2008). The Bayesian lasso. *Journal of the American Statistical Association*, *103*(482), 681–686. https://doi.org/10.1198/016214508000000337

Piironen, J., & Vehtari, A. (2017a). Comparison of Bayesian predictive methods for model selection. *Statistics and Computing*, *27*(3), 711–735. https://doi.org/10.1007/s11222-016-9649-y

Piironen, J., & Vehtari, A. (2017b). Sparsity information and regularization in the horseshoe and other shrinkage priors. *Electronic Journal of Statistics*, *11*(2), 5018–5051. https://doi.org/10.1214/17-ejs1337si

Requia, W. J., Adams, M. D., Arain, A., Papatheodorou, S., Koutrakis, P., & Mahmoud, M. (2018). Global Association of Air Pollution and Cardiorespiratory Diseases: A Systematic Review, Meta-Analysis, and Investigation of Modifier Variables. *American Journal of Public Health*, *108*(S2), S123–S130. https://doi.org/10.2105/AJPH.2017.303839

Riley, R. D., Higgins, J. P. T., & Deeks, J. J. (2011). Interpretation of random effects meta-analyses. *BMJ*, *342*, d549. https://doi.org/10.1136/bmj.d549

Rosettie, K. L., Joffe, J. N., Sparks, G. W., Aravkin, A., Chen, S., Compton, K., … Murray, C. J. L. (2021). Cost-effectiveness of HPV vaccination in 195 countries: A meta-regression analysis. *PLOS ONE*, *16*(12), e0260808. https://doi.org/10.1371/journal.pone.0260808

Röver, C., Bender, R., Dias, S., Schmid, C. H., Schmidli, H., Sturtz, S., … Friede, T. (2021). On weakly informative prior distributions for the heterogeneity parameter in Bayesian random-effects meta-analysis. *Research Synthesis Methods*, *12*(4), 448–474. https://doi.org/10.1002/jrsm.1475

Schooler, J. (2011). Unpublished results hide the decline effect. *Nature*, *470*(7335), 437–437. https://doi.org/10.1038/470437a

Sebri, M., & Dachraoui, H. (2021). Natural resources and income inequality: A meta-analytic review. *Resources Policy*, *74*, 102315. https://doi.org/10.1016/j.resourpol.2021.102315

Stan Development Team. (2022). *RStan: The R interface to Stan*.

Thompson, C. G., & Becker, B. J. (2020). A group-specific prior distribution for effect-size heterogeneity in meta-analysis. *Behavior Research Methods*, *52*(5), 2020–2030. https://doi.org/10.3758/s13428-020-01382-8

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

Tibshirani, R. (2011). Regression shrinkage and selection via the lasso: A retrospective. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *73*(3), 273–282. https://doi.org/10.1111/j.1467-9868.2011.00771.x

van Erp, S., Oberski, D. L., & Mulder, J. (2019). Shrinkage priors for Bayesian penalized regression. *Journal of Mathematical Psychology*, *89*, 31–50. https://doi.org/10.1016/j.jmp.2018.12.004

Zhang, P. (1992). Inference after variable selection in linear regression models. *Biometrika*, *79*(4), 741–746. https://doi.org/10.1093/biomet/79.4.741