Author's response to reviews

Title: Risk of bias: Power to detect study level moderator effects in meta-analysis

Authors:

   Susanne Hempel (susanne_hempel@rand.org)
   Jeremy N.V. Miles (jeremy_miles@rand.org)
   Marika J. Booth (suttorp@rand.org)
   Zhen Wang (wang.zhen@mayo.edu)
   Sally C. Morton (scmorton@pitt.edu)
   Paul G. Shekelle (paul_shekelle@rand.org)

Version: 2 Date: 30 July 2013

Author's response to reviews: see over
Dear Dr. Rodgers,

We would like to resubmit our article ‘Risk of bias: Power to detect study level moderator effects in meta-analysis’.
We thank the editors and the reviewer for the detailed and constructive comments. Please find below how we have addressed them.

Yours sincerely,

Susanne Hempel
(on behalf of all authors)

Response to comments

Reviewer: Arianne Verhagen

General comments
1. For me it is not clear what the objective is. In the abstract is stated: “. . .to investigate the effect of trial size, . . .on power to detect moderator effects . “. Later on in the manuscript it seems that the authors are only evaluating trial quality as a possible moderator, using the other variables as factor to simulate with (see conclusion, this only refers to study quality). If that is the case. Please state clearly and also you might consider changing the title also for clarity.

Thank you for this observation. We have used study quality as a typical example of a study-level moderator, a moderator that readers are most familiar with. However, the analyses hold for any study level moderator. We have added an explanation to the implication section of the discussion (second sentence) to make this more explicit.

2. Throughout the manuscript the authors use the term 'quality of studies', meaning the methodological quality. I prefer the term 'risk of bias' instead because this term more clearly grasps what is meant with methodological quality.

In this case we think it is important to differentiate ‘quality’ (the study characteristic) and risk of ‘bias’ (the potential effect on the treatment estimate). We have added a reference to bias to the discussion section to address this point (implication section in the discussion).

3. Not clearly stated, but I assume that their null-hypothesis is that there is no moderator effect of trial quality on effect sizes. Their main conclusion is that due to heterogeneity this null-hypothesis cannot easily be rejected and this explains, according to the authors, the conflicting results from previous studies. It seems that the authors are convinced that the truth is that the null-hypothesis should be rejected. Given a dataset with hardly any heterogeneity and sufficient power one can be able to significantly show moderator effect of trial quality. Of course, given these preconditions one can get everything and anything significant. It might also be possible that whenever there is a moderating effect of trial quality on effect sizes it might depend on other factors like the topic of research. Trial quality might not be a generic threat to true outcomes.

Our analyses show that most detecting moderator effects requires more powerful analyses than are employed in most published investigations, hence negative findings should not be considered evidence of lack of effects and investigations are not hypothesis-proving unless a power calculation shows sufficient ability to detect effects. In the simulations we know that the moderator has an effect on the effect sizes because we set up the simulations accordingly. However, the analyses show that only under special circumstances can this (definitely present)
effect be detected in meta-regressions. We have revised the respective paragraph in the
implication section of the discussion to clarify the issue.

4. This manuscript is very lengthy and needs shortening and a more concise writing.
We have shortened the manuscript, in particular the background section, to address this point.

Specific comments
1. Abstract. Please state the objective clearly in the abstract. Now I think the objective is in the method
   section of the abstract?
We have revised the abstract accordingly and have added an explicit sentence stating the
objective to the background section of the abstract.

2. Abstract, results. Please delete ‘only’ in the second sentence, it is used twice in one sentence.
Changed accordingly

3. Background. This section is rather long and needs to be shortened. For instance, the second half of page
   5 and the first half of page 6 all try to explain that research on the relationship of risk of bias on study
   outcome, or outcome of meta-analysis is conflicting. Please shorten the background.
Changed accordingly

4. Background, page 6, last lines. The authors describe the difference between subjective and objective
   outcomes. This is an irrelevant difference. Outcome measurement and outcome measures should be valid,
   and according to me relevant to the patient (patient reported outcomes). The main difference the authors are
   pointing out is that an outcome such as ‘mortality’ does not suffer from bias due to a lack of blinding as
   most patient reported outcome measures do. Please discuss the lack of blinding as a factor related to
   ‘exaggerated’ effect sizes.
The paragraph reports the results found by Wood as a rare example of a study investigating
variables that determine when associations between study characteristics and effect sizes are
detected. Rather than getting deeper into a discussion of subjective versus objective outcomes we
have shortened the paragraph.

5. Background, page 7 last para. Here the authors mix terms like ‘between cluster (study) variance’ and
   ‘intra-cluster (or intra-class) correlation’. These terms are confusing. Intra-cluster (or intra-class)
correlation refers to test-retest reliability and ICC as a measure of agreement. I would like the authors to
refrain from these terms, they are not necessary for the rest of the study, so please delete them. Also, the
sentence: ‘greater variance between studies is associated with a larger intra-study correlation’ is
problematic, because this is not necessarily so. Please shorten this whole para, you just try to explain what
heterogeneity is, which can be done much more concise.
We were attempting to relate the power of meta-analyses to established knowledge of power in
multilevel models, but we accept that this is unclear without a great deal of addition elucidation
and have therefore removed the reference to multilevel models.

6. Methods, first para. Here the authors define that they differentiate the studies into high and low quality.
Besides the fact that I would prefer the authors to differentiate between high, unclear and low risk of bias,
there is no mention of cut-off points. When do authors consider a study of high quality? Please define.
Please also explain the rationale for choosing continuous outcome measures only.
Given the complexity of the analysis we have chosen a dichotomous predictor and continuous
outcome measures. A concrete translation of a dichotomous predictor is the differentiation into
high quality and low quality studies. In our experience readers will familiar with such as distinct
categorization used for the models and discussing thresholds and potential definitions of what
quality could mean in empirical studies would be too confusing. An analysis of categorical
outcomes can be associated with other methodological issues, such as continuity correction, and the choice of appropriate transformation of the probabilities, and this would be a separate paper.

7. Methods, page 9, first para. Please explain why you choose for the Tau as a measure of heterogeneity. The I2 is much more often used and easier to interpret. We have discussed the choice of heterogeneity measure extensively. We had originally used I-square for our simulation work but have been convinced by peer reviews received for the AHRQ methods report on Detection of Associations Between Trial Quality and Effect Sizes that the characteristics of tau-squared is more suitable to model heterogeneity. While I-squared describes the percentage of total variation across studies that is due to heterogeneity rather than chance, tau-squared describes the residual variance in the effect sizes. If additional trials are added to a meta-analysis, and these trials are exactly the same as the trials that were already in the meta-analysis, the value of tau-squared will not change, the value of I-squared will change. We have added a sentence explaining the characteristics to the Simulation Design section of the Method section.

8. Please split the results from the discussion sections, and please shorten the discussion part. The combined result and discussion section format is based on the journal’s guidance for authors. We have considered other formats but we feel that separating the discussion from the result section might be too difficult to follow and would potentially add to the length of the manuscript, giving the complexity of the material.

**Reviewer: Mark Simmonds**

**Major compulsory revisions.**

1. Contrary to your claim at the end of the background section there is considerable literature around the limited power and possible bias is estimating moderator effects in meta-analysis (papers by for example, Thompson, Higgins, Berlin). Key among these papers is that the power to detect moderator effects has already been determined algebraically in the absence of heterogeneity (Simmonds & Higgins, Stat. Med. 2007). The findings of that paper have a substantial impact on your work, which you must consider and properly acknowledge.

We thank the reviewer for this suggestion and have added references to patient-level moderator research. In addition, we have explained more clearly that our paper is about study-level moderator effects.

2. (Background) This section is excessively long. Do not lecture the reader on the history or importance of randomisation, quality assessment or statistical power. This section should cover only issues in meta-analysis particularly important for this paper.

We have shortened the background section accordingly.

3. The background section focuses on moderator effects that may cause bias (eg low quality), but there are many other “positive” moderator effects (eg. dose-response). Are you interested only in bias-causing effects or all moderator effects? The whole paper needs more clarity on this issue.

We concur; please see our response to comment 1 made by reviewer 1.

4. (Methods) How are moderator effects estimated? Using meta-regression or subgroup analysis? The method must be described as it will affect the power. Generally the statistical analysis methods should be presented in more detail.

We have added to the description of the analysis, and explained that a meta-regression was estimated.
5. (Application examples) I found this section confusing. Are you just describing some practical examples here? If so why the discussion of Monte Carlo simulations? Please clarify this section. Perhaps you could summarise your examples in a table.

We have added a short introduction to the application examples in order to clarify the purpose of these datasets in the analytic approach.

6. Given the existence of the paper mentioned above the simulation when there is no heterogeneity is unnecessary and would be better replaced with the correct theoretical power from that paper. A simulation is only needed when there is residual heterogeneity where there is (probably) no algebraic form for the power.

We agree that a simulation is not necessary with there is no residual heterogeneity. However, we feel that the confidence of the reader in the simulations is improved with inclusion of these results, which agree with that which could be calculated, hence we have retained them.

7. The paper mentioned above also shows that the power is a function of the relative sample size in the two groups of studies (or the total sample size for a 50:50 split) so it seems that presenting results in terms of number of trials and sample size in each trial is unnecessary and your figures could be simplified. The "covariate heterogeneity" is a key determinant of power that you should consider. In particular you should present figures for other splits of the data, not just 50:50 and 25:75, noting that it is the total sample size in each group, not the number of trials, that drives the power.

We are not sure what the reviewer is asking here - the two different splits of the data (50:50 and 75:25) were chosen as illustrative examples.

8. I assume you have used a 5% Type I error rate calculate power, but if we are looking for bias a much higher type I error rate would be acceptable, to avoid not spotting bias. Higher error rate means greater power. This should be discussed.

Added as suggested.

Minor essential revisions

9. (Applications, 1st parag.) Researchers are surely not interested in how many studies are needed to detect a moderator effect. First because it is sample size, not number of studies, that is important, and second because in a meta-analysis we have no control over the number of studies. This should be removed or clarified.

We have revised the sentence but we strongly believe that power analyses for meta-regressions should be considered. In the meta-epidemiological investigations we are looking at, the limiting factor is often not the number of trials that could be included, but the resources available to read and categorize these studies.

10. Potential for bias is a major issue in estimating moderator effects in meta-analysis. This should be discussed in the background or discussion section.

We have added a reference to bias; see our response to comment 2 made by reviewer 1.

Discretionary revisions

11. (Methods) I am concerned that a moderator effect of only 0.4 at most is too small. If an effect is small it may not be clinically or practically meaningful. Larger effects are of greater interest. If possible give results for larger effect sizes (e.g. 1 SD).

We respectfully disagree with the reviewer. The effect size that we present is the difference in effect sizes across studies with high and low quality. Conventionally, effect sizes are given as 0.2 for small, 0.5 and 0.8 for large. The largest effect we consider, of 0.2, is almost sufficient to
‘shift’ the effect size from a moderate to a small effect, and is broadly equivalent to the effects that were found in the published research.

12. (Methods) Heterogeneity is better quantified using $I^2$ rather than $tau^2$ for ease of interpretation and to maintain heterogeneity levels as numbers of trials change. If possible perform the simulation in terms of $I^2$.

Please see our response to reviewer 1, comment 7.

**Reviewer: Joseph Lau**

The paper by Hempel et al. in essence asks the question of whether many of the current meta-epidemiological studies to identify potential moderator effects (study quality features) are sufficiently powered to take on this task. The authors constructed a Monte Carlo simulation study using a wide range of parameters as well as on a cross section of 200 randomly selected RCTs published in 2006. In addition, the authors performed post-hoc power calculations on four previously published meta-epidemiologic datasets. The conclusion that “negative findings (from meta-epidemiological research) should not be considered evidence of lack of effects” is not surprising and I doubt that most researchers in this area would think otherwise. Even (hypothetically) if meta-epidemiological studies show no significant difference between randomized studies and non-randomized studies, I would not conclude that therefore randomization is unnecessary. Thus, I see this paper mainly contributes to the methodological discussion by illustrating how difficult (infeasible or perhaps nonsensical) it is to carry out this type of meta-epidemiological research. To be able to identify potential moderator effect, one needs a very large number of trials that is not present in the current published research on this topic. The number of trials needed further increases in the presence of heterogeneity.

While the simulation covered a wide range of values (number of subjects in trials, number of trials in meta-analyses, different $tau$-squared values, and effect sizes), it did not take into the account of potential different magnitude of moderator effects in heterogeneous clinical settings of the meta-analyses. For example, Schultz et al reported in JAMA in 1995 after examining 33 obstetrics meta-analyses that trials with inadequate method of treatment allocation resulted in 41% large effects, trials with unclear method of treatment allocation have 30% larger effects, and trials not double blind have 17% larger effects. As you noted, these observations were not consistently observed by others examining similar questions across other medical domains. While blinding could be readily achieved in drug trials, it is unachievable most of the time in surgical trials. Similarly there are differences of moderator effects for objective versus subjective outcomes. Further, even if this observation in a cross-section of time is true, could it change over time?

Thus, moderator effect may vary by clinical conditions, the type of outcomes reported, and other factors. While some of the moderators may have the same effect across all clinical settings, some may be quite different. Modeling this heterogeneity into the simulation will likely require more trials and meta-analyses for meta-epidemiological studies.

My take-home message from this simulation study is that to demonstrate a moderator effect using the current approach of meta-epidemiological studies, one needs a very large number (perhaps infeasibly so) of trials from many meta-analyses. Even if we are able to have adequately powered meta-epidemiological studies in the future, will these results be useful to inform systematic review practice? The results of these analyses are findings of various degrees of association and/or ratio of odds ratio and the like. The challenge of using this information is given a specific meta-analysis, how would one apply and interpret moderator effects from meta-epidemiological studies? Unless we know for certain that a specific factor is truly universal across all clinical conditions, study designs, etc. It would be difficult to interpret and apply these results. Should one be adjusting the effect size of meta-analysis using meta-epidemiological studies?

We thank the reviewer for the insightful comments although we respectfully disagree with the view that the lack of power for meta-regressions is generally known and accepted. Meta-epidemiological studies continue to be published and are frequently cited and critical appraisal measures are frequently chosen based on their empirical evidence of bias.