## PEER REVIEW HISTORY



## ARTICLE DETAILS

| | |
|---|---|
| **TITLE (PROVISIONAL)** | Effects of experimental sleep deprivation on aggressive, sexual and maternal behavior in animals – A systematic review protocol |
| **AUTHORS** | **Gabriel Natan Pires (Corresponding Author)**<br>**Andréia Gomes Bezerra**<br>**Rob B.M. de Vries**<br>**Cathalijn H.C. Leenaars**<br>**Merel Ritskes-Hoitinga**<br>**Sergio Tufik**<br>**Monica Levy Andersen** |

## VERSION 1 - REVIEW

| | |
|---|---|
| **REVIEWER 1** | ***Olavo Amaral***<br>***Instituto de Bioquímica Médica***<br>***Please state any competing interests or state 'None declared':*** ***None declared*** |
| **REVIEW RETURNED** | 23-02-18 |

| | |
|---|---|
| **GENERAL COMMENTS** | Major points:<br><br>Introduction/Discussion:<br><br>- Both the introduction and discussion mention that the meta-analysis is important to explore the "animal-to-human translational potential of animal studies in this field" and to optimize experimental design to increase translational reliability. While I agree with this goal, the authors never mention what is known about the effects of sleep deprivation in humans on the three dimensions of behavior to be analyzed (aggressive, sexual and maternal behaviors). If comparing animal data to human data is among their main objectives, the best available evidence on this subject should be at least touched upon in the protocol, so that the reader will know to what human results the authors will be comparing their data with.<br><br>Data analysis:<br><br>- As in most meta-analyses in basic science, the authors will probably face the problem of non-independence among experimental results – as, although experiments are theoretically independent, it is very likely that results from the same paper (or from the same research group, for that matter) will be more similar among themselves than when compared to those of other |

| | |
|---|---|
| | papers/groups, therefore. This problem can be glimpsed clearly from the authors' previous meta-analysis (Pires et al., 2016), in which many papers have multiple results (up to 6 in some cases), and in which results from the same article/group are usually similar among themselves (see, for example, the articles from Kalonia and Kumar, Kumar and Kalonia and Kumar and Singh, whose effect sizes markedly differ from the rest of the sample. As an excessive number of non-independent results from the same article/research group can sometimes lead this single article/group to reach a larger weight in the meta-analysis, and thus bias effect estimates obtained for random effects models, an interesting alternative would be to use multilevel modeling in order to control for this effect. Using the article as a level in the model is rather straightforward, and has been done in other meta-analysis of animal experiments in neuroscience (see for example Kredlow et al., Psychol Bull 2016; 142:314-336). Using research group as a level is less straightforward, as it defines an operational definition of research group. Nevertheless, we are currently working on a tool to define author clusters based on collaboration graphs, and would be happy to share it with the authors once it is sufficiently developed.<br><br>- In evaluating publication bias, why don't authors go beyond visual analysis and actually evaluate it through a formal procedure (e.g. Egger's regression, excess significance tests), as well as correct for it if necessary using trim-and-fill analysis? These are rather standard procedures and would certainly strengthen the evidence concerning publication bias in this case.<br>- Stratified subgroup analysis seems like a good option to investigate the impact of categorical variables (e.g. species, sex, type of deprivation) on effect sizes. However, for quantitative variables (e.g. age, length of sleep deprivation), meta-regression looks like a better option, as it will avoid unnecessary discretizing of the data.<br><br>- An interesting potential benefit of meta-analysis is that, once a best estimate of the true effect size in a given field is obtained, it can be used to calculate statistical power of the individual studies included (for an example, see Button et al., Nat Rev Neurosci 2013; 14: 365-376), which can provide interesting insights on whether sample sizes commonly used in the field are adequate or not. If the authors agree with the utility of this, they might consider including this procedure in the protocol. It is important to note that this only makes sense if a significant effect of the tested interventions is indeed found in the meta-analysis (see Nord et al., J Neurosci 2017; 37: 8051-8061).<br><br>Minor points:<br><br>Abstract: |

| | |
|---|---|
| | - It is not completely clear in the abstract that the three systematic reviews relate to aggressive, sexual and maternal behavior respectively (as the 3 types of behavior are mentioned only in the first sentence). Although most readers will probably guess this, the point could be made clearer.<br><br>- "Visual analysis of funnel plots to evaluate publication bias" is a fragment (i.e. there's a verb missing).<br><br>Introduction:<br><br>- "Comorbidity" is much more usual than "co-morbidity" in the literature.<br><br>- "Sleep deprivation being able" to do something sounds like a weird formulation for a sentence.<br><br>- "Several animal models of sleep deprivation and several behavioral… " – One probably does not need to repeat "several" in this sentence.<br><br>- "Motivated social behaviors, such as, aggressive…" – There should be no comma after "such as"<br><br>- "Sleep-deprived" should be hyphenized.<br><br>Methods:<br><br>Bibliographic Search:<br><br>- All strings for sleep deprivation are listed as "sleep _____", and thus might be conceivably sensitive to the inversion of word order (e.g. "the effects of restricting sleep in rodents"). Although it is unlikely that a study would not include at least one of the included terms, the authors could consider using "sleep" AND ("deprivation" OR "restriction" OR etc.), or at least include some terms in reversed order (e.g. "loss of sleep"). I'd additionally add "sleep-deprived" and "sleep-restricted", in case these are not already included in the MeSH terms.<br><br>- Similarly, for aggressive behavior, I think it would also make sense to include "violent", if that is not covered by MeSH terms.<br><br>- For maternal behavior, couldn't the authors substitute the various forms of "maternal behavior" (e.g. "maternal behavio*", "maternal care", etc.) simply by the word "maternal" (or "maternal*")?<br><br>Study selection: |

| | |
|---|---|
| | - It is not clear whether discrepancies will be solved by discussion and consensus in both phases (abstract and full-text screening) or only for full-text.<br><br>- “lists of articles included” should be “lists of included articles”.<br><br>Inclusion and exclusion criteria:<br><br>- “full text are not available” should be “full text is not available”.<br><br>- “Not an experimental type of article” could be simplified as “not an experimental article”.<br><br>- “No presence of control group…” could be simplified as “Absence of control group…” or “No control group…”.<br><br>Data extraction:<br><br>- For the intervention characteristics, if studies with partial sleep restriction are to be included, it is unclear what “duration of sleep deprivation” means, as this could mean both (a) the duration for which animals are restricted over each 24-hour cycle and (b) the total duration of sleep restriction (which could happen over several days). Thus, the authors should consider using these as separate variables (or, alternatively, include a “time of onset” and “time of offset” so the duration could be calculated).<br><br>- Common problems for data extraction in meta-analysis include error bars in articles not being described as standard deviations or standard error, or sample size being given as a range. In these cases, one can opt either for conservative inclusion of the data (i.e. assuming that data is S.D. or the lower number of the range) or exclusion. From what was stated in this section (i.e. “If an article does not fully describe the data...), I got the feeling that the authors mean to opt for exclusion; nevertheless, the data analysis section later states that, for sample size ranges, results will be included considering the lowest possible value. It is unclear, though, if a similar “conservative inclusion” procedure will be done in the case of undefined error bars.<br><br>Risk of bias and publication bias:<br><br>- I disagree with the affirmation that the studies cannot be blinded due to housing conditions. Although this makes blinding require two separate experimenters during behavioral sessions (one for removing the animals for housing, the other for assessing behavior), it is certainly feasible. Sleep-deprived animals could theoretically be recognizable if the effects of deprivation as marked, but that could be claimed for pretty much any intervention that affects behavior – there is nothing special about sleep deprivation in this sense. Thus, I would recommend |

| | |
|---|---|
| | including blinding in the risk of bias tool.<br><br>Data Synthesis:<br><br>- "Whenever all experiments in an analysis use the same behavioral test and<br>the same unit of measurement, and follow the same experimental standards in data acquisition, the effect size within our comparison will be a mean difference."<br>This is somewhat vague – what exactly do the authors mean by "same experimental standards"?<br><br>- This may be a matter of personal preference, but why not use forest plots for stratified analyses as well? The bar graphs (as presented in Pires at al., 2016) do not convey an idea of sample size or precision of individual studies, and thus present much less information than the forest plots themselves.<br><br>Discussion:<br><br>- "the effects observed in animals is…". Should be "are".<br><br>- "Showing that lack of sleep leads to a decrease…". This sentence is a fragment.<br><br>- "Both spontaneous and experimental sleep deprivation always coincides…". Should be "coincide".<br><br>- The "Profile of scientific output in the field" seems like an interesting benefit of the study, but the authors never mentioned in the analysis section of the protocol that they would analyze the results by geographical region or year of publication. If they indeed plan to do this, it should be included in the data analysis section. |

| | |
|---|---|
| **REVIEWER 1** | ***Malcolm Macleod***<br>***University of Edinburgh***<br>***Please state any competing interests or state 'None declared': I collaborrate with the SYRCLE group*** |
| **REVIEW RETURNED** | 27-02-18 |

| | |
|---|---|
| **GENERAL COMMENTS** | Thank you for the opportunity to review this systematic review protocol. I have some comments:<br>1. The Syrcle group have themselves shown the limited utility of funnel plotting in the context of SMD analysis where the weighting is based on variance. This needs to be better justified.<br>2. Avoidance of the non- peer reviewed literature is questionable, especially if you expect to see publication bias, and given the emerging use of BioRxiv by the community. |

| | |
|---|---|
| | 3. The authors may wish to reflect on the statistical power of partitioning heterogeneity in SMD analyses in the context of Wang et al BioRxiv 2018 ... unless they have large numbers of studies, they are likely to have limited power. It is what it is, but would be reasonable to discuss.<br>4. There could be more definite statement of the stratifications, and what "where possible" means, and what the critical alpha will be for these comparisons (ie are you going to correct for 7 contrasts.<br>5. I half - get the difficulty in blinding outcome assessment, but there are still ways in which some of the experiment (data analysis, exclusion of outliers etc) might be blinded. And I dont understand why randomisation isn't a thing. And ref 27 is a review of human studies which includes an item on baseline sleep habits ... and is psuedoreplication not an issue in this space, if sleep deprivation exposures are at the level of the cage? So I would articulate, here, the risk of bias items, rather than having the reader track back |

## VERSION 1 – AUTHOR RESPONSE

Reviewer 1:
Major points:
Introduction/Discussion:
- Both the introduction and discussion mention that the meta-analysis is important to explore the “animal-to-human translational potential of animal studies in this field” and to optimize experimental design to increase translational reliability. While I agree with this goal, the authors never mention what is known about the effects of sleep deprivation in humans on the three dimensions of behavior to be analyzed (aggressive, sexual and maternal behaviors). If comparing animal data to human data is among their main objectives, the best available evidence on this subject should be at least touched upon in the protocol, so that the reader will know to what human results the authors will be comparing their data with.
This is indeed and important consideration. A brief yet well referenced sentence was added to the introduction, in order to shortly denote the results of sleep deprivation on aggressive, sexual and maternal behaviors in humans (page 4).

Data analysis:
- As in most meta-analyses in basic science, the authors will probably face the problem of non-independence among experimental results – as, although experiments are theoretically independent, it is very likely that results from the same paper (or from the same research group, for that matter) will be more similar among themselves than when compared to those of other papers/groups, therefore. This problem can be glimpsed clearly from the authors’ previous meta-analysis (Pires et al., 2016), in which many papers have multiple results (up to 6 in some cases), and in which results from the same article/group are usually similar among themselves (see, for example, the articles from Kalonia and Kumar, Kumar and Kalonia and Kumar and Singh, whose effect sizes markedly differ from the rest of the sample. As an excessive number of non-independent results from the same article/research group can sometimes lead this single article/group to reach a larger weight in the meta-analysis, and thus bias effect estimates obtained for random effects models, an interesting alternative would be to use multilevel modeling in order to control for this effect. Using

the article as a level in the model is rather straightforward, and has been done in other meta-analysis of animal experiments in neuroscience (see for example Kredlow et al., Psychol Bull 2016; 142:314-336). Using research group as a level is less straightforward, as it defines an operational definition of research group. Nevertheless, we are currently working on a tool to define author clusters based on collaboration graphs, and would be happy to share it with the authors once it is sufficiently developed.
This is a very interesting topic and I am glad the Reviewer has checked our previous meta-analysis to address it. The results of Kumar's group seems more as an unadvised data recycling, in which control and naïve groups data were reused in all the eight articles, rather than a case of consistent data replication. A deeper data checking shows that some results are absolutely identical in different articles. This was a flag for our suspicion of data recycling, reason why we performed a sensitivity analysis in that case. Since we are not qualified to judge the ethics and procedures behind this potential data recycling, we thought it would be more careful to perform a sensitivity analysis on that case.
That said, I think using Kumar's data may not be the best proof of a "research group bias". Conversely, analyzing the consistency on the results of other research groups shows some cases of intra-group inconsistency. Maybe the biggest proof can be seen on the research group of Dr. Monica Andersen (to which four authors in this manuscript belongs to), which has contributed with seven articles in our meta-analysis, presenting standardized effect sizes ranging from -3.79 to 2.60.
In any case, we do agree that we can have some sort of "research group bias" in most cases. I would appreciate to know more about the methods this Reviewer is working on, but while it is not fully developed, my suggestion will be to evaluate it by means of both stratified and sensitivity analysis. We will perform stratified analysis specific to each research group whenever it has at least three articles included in our sample, as well as sensitivity analysis excluding a given research group from the pool whenever it contributed with at least 10% of the total sample. These suggestions have been included into the methods section (page 12).

- In evaluating publication bias, why don't authors go beyond visual analysis and actually evaluate it through a formal procedure (e.g. Egger's regression, excess significance tests), as well as correct for it if necessary using trim-and-fill analysis? These are rather standard procedures and would certainly strengthen the evidence concerning publication bias in this case.
We appreciate the Reviewer's suggestion and agree that both trim and fill analysis and Egger's regression would provide important additional information. Both were included into our protocol (page 11).

- Stratified subgroup analysis seems like a good option to investigate the impact of categorical variables (e.g. species, sex, type of deprivation) on effect sizes. However, for quantitative variables (e.g. age, length of sleep deprivation), meta-regression looks like a better option, as it will avoid unnecessary discretizing of the data.
While we do agree that meta-regressions are a better option to numeric variables, we prefer to stick with the stratified analysis in our protocol for two reasons: 1. Meta-regression are better to provide precise estimates of effects, but stratified analysis are a good option to analyze the effects of potential confounders and to check the consistency of the data in more specific cases. 2. Our most important quantitative variable (length of sleep deprivation) is kind of categorical in nature, since in most of the experiments (especially those working with sleep restriction and REM sleep deprivation protocols) the length of the protocols is a multiple of 24h.

- An interesting potential benefit of meta-analysis is that, once a best estimate of the true effect size in a given field is obtained, it can be used to calculate statistical power of the individual studies included (for an example, see Button et al., Nat Rev Neurosci 2013; 14: 365-376), which can provide interesting insights on whether sample sizes commonly used in the field are adequate or not. If the

authors agree with the utility of this, they might consider including this procedure in the protocol. It is important to note that this only makes sense if a significant effect of the tested interventions is indeed found in the meta-analysis (see Nord et al., J Neurosci 2017; 37: 8051-8061).
This is a very interesting approach, of which we were not aware. It certainly could contribute to the overall field of basic behavioral research. However, stating how big sample sizes should be goes beyond our current research aims. We opt not to include it in the current protocol, but we are open to consider it as independent experiments in the future, and may suggest this in the discussion of the resulting review.

Minor points:
Abstract: - It is not completely clear in the abstract that the three systematic reviews relate to aggressive, sexual and maternal behavior respectively (as the 3 types of behavior are mentioned only in the first sentence). Although most readers will probably guess this, the point could be made clearer.
The abstract was rewritten, in order to make the meta-analysis distribution clearer, as suggested (page 2).

- "Visual analysis of funnel plots to evaluate publication bias" is a fragment (i.e. there's a verb missing).
Changed, as suggested.

Introduction: - "Comorbidity" is much more usual than "co-morbidity" in the literature.
Changed, as suggested.

- "Sleep deprivation being able" to do something sounds like a weird formulation for a sentence.
Changed, as suggested

- "Several animal models of sleep deprivation and several behavioral… " – One probably does not need to repeat "several" in this sentence.
Changed, as suggested

- "Motivated social behaviors, such as, aggressive…" – There should be no comma after "such as"
Changed, as suggest

- "Sleep-deprived" should be hyphenized.
Changed throughout the text, as suggested

Methods:

Bibliographic Search:
- All strings for sleep deprivation are listed as "sleep _____", and thus might be conceivably sensitive to the inversion of word order (e.g. "the effects of restricting sleep in rodents"). Although it is unlikely that a study would not include at least one of the included terms, the authors could consider using "sleep" AND ("deprivation" OR "restriction" OR etc.), or at least include some terms in reversed order (e.g. "loss of sleep"). I'd additionally add "sleep-deprived" and "sleep-restricted", in case these are not already included in the MeSH terms.
We have updated the search string for sleep deprivation, as suggested. On a preliminary test, it returns twice the number of results in comparison to the previous search string (page 6). The revised one will be
sleep deprivation[mesh] OR ((sleep[tiab] OR REM[tiab]) AND (depriv*[tiab] OR restrict*[tiab] OR fragment*[tiab] OR curtailment[tiab] OR loss[tiab] OR disrupt*[tiab] OR disturb*[tiab]))

- Similarly, for aggressive behavior, I think it would also make sense to include “violent”, if that is not covered by MeSH terms.
We included violen*[tiab] as a complement to the previous search string, as suggested (page 7).

- For maternal behavior, couldn’t the authors substitute the various forms of “maternal behavior” (e.g. “maternal behavio*”, “maternal care”, etc.) simply by the word “maternal” (or “maternal*”)? Including only “maternal” instead of more specific terms will result in a substantial amount of false negative search results. In a previous test, it increased the number of records with about 30%. Since “maternal” can be related to a broad selection of terms completely unrelated to maternal behavior (eg.: maternal health, maternal-fetal, maternal exposure, maternal deprivation, etc.) we preferred to stick to a more specific search string.

Study selection: - It is not clear whether discrepancies will be solved by discussion and consensus in both phases (abstract and full-text screening) or only for full-text.
Now it is mentioned that discussion and consensus will be used to solve disputes in both selection phases (page 7).

- “lists of articles included” should be “lists of included articles”.
Changed, as suggested

Inclusion and exclusion criteria: - “full text are not available” should be “full text is not available”.
Changed, as suggested

- “Not an experimental type of article” could be simplified as “not an experimental article”.
Changed, as suggested

- “No presence of control group…” could be simplified as “Absence of control group…” or “No control group…”.
Changed, as suggested

Data extraction:
- For the intervention characteristics, if studies with partial sleep restriction are to be included, it is unclear what “duration of sleep deprivation” means, as this could mean both (a) the duration for which animals are restricted over each 24-hour cycle and (b) the total duration of sleep restriction (which could happen over several days). Thus, the authors should consider using these as separate variables (or, alternatively, include a “time of onset” and “time of offset” so the duration could be calculated).
This is indeed an important consideration. In order to provide more detailed data in cases of sleep restriction protocols, we have specified that “duration of sleep deprivation” refers to the total duration of the experimental protocol. We have also inserted “daily sleep opportunity onset” and “duration of daily sleep opportunity onset” for cases of sleep deprivation (page 10).

- Common problems for data extraction in meta-analysis include error bars in articles not being described as standard deviations or standard error, or sample size being given as a range. In these cases, one can opt either for conservative inclusion of the data (i.e. assuming that data is S.D. or the lower number of the range) or exclusion. From what was stated in this section (i.e. “If an article does not fully describe the data...), I got the feeling that the authors mean to opt for exclusion; nevertheless, the data analysis section later states that, for sample size ranges, results will be included considering the lowest possible value. It is unclear, though, if a similar “conservative inclusion” procedure will be done in the case of undefined error bars.

This is an interesting consideration. The resolution of cases of imprecise sample range was already expected, and our strategy is to consider the lower possible value on a given sample. For cases in which errors bars are not described as SD or SEM, we will approach as follows: first we will try to contact the authors. If that is not possible, we will then assume the error bars as a measure of SD (page 10).

Risk of bias and publication bias: - I disagree with the affirmation that the studies cannot be blinded due to housing conditions. Although this makes blinding require two separate experimenters during behavioral sessions (one for removing the animals for housing, the other for assessing behavior), it is certainly feasible. Sleep-deprived animals could theoretically be recognizable if the effects of deprivation as marked, but that could be claimed for pretty much any intervention that affects behavior – there is nothing special about sleep deprivation in this sense. Thus, I would recommend including blinding in the risk of bias tool.
There are two blinding-related items in the risk of bias assessment: blinding during the experiment (performance bias) and blinding during outcome detection (detection bias). According to the SYRCLE risk of bias tool, blinding as a performance bias is assessed by the following question: Were the caregivers and/or investigators blinded from knowledge which intervention each animal received during the experiment? This kind of blinding is methodologically impossible because any caregiver handling the animals will immediately recognize those housed in standard conditions and those housed in sleep deprivation apparatuses. Blinding in outcome detection is assessed by the following question: "Was the outcome assessor blinded? In this case, despite we still think any sleep deprivation can be easily recognized by animals body characteristics (posture, sleep pressure, shut eyes, fur, etc), we do acknowledge it is methodologically possible to blind outcome assessors. Thus, we prefer to stick with the decision to remove the blinding on performance out of our list. However, blinding in outcome assessment will be incorporated back in our risk of bias analysis (page 11).

Data Synthesis: - "Whenever all experiments in an analysis use the same behavioral test and the same unit of measurement, and follow the same experimental standards in data acquisition, the effect size within our comparison will be a mean difference." This is somewhat vague – what exactly do the authors mean by "same experimental standards"?
By "experimental standards" we meant to say the same animal model characteristics, intervention characteristics and outcome measures, as depicted in table 2. If these data are used consistently and equally among two or more studies, we believe it is feasible to use simple mean difference as a matter of effect size. The sentence was rewritten to make it clearer, in order to denote that this decision will be made based on practical parameters, rather than on vague intuition (page 11).

- This may be a matter of personal preference, but why not use forest plots for stratified analyses as well? The bar graphs (as presented in Pires at al., 2016) do not convey an idea of sample size or precision of individual studies, and thus present much less information than the forest plots themselves.
Both types of graphical representation have their strengths and we believe we can use either in different contexts. The "primary" meta-analysis is better disclosed by means of forest plots, cases in which the contribution of each article to the overall estimate is an important information. In cases of stratified analysis, more than only providing the overall estimate, it is interesting to visually analyze which cases get closer to the overall estimate as well as to know which contributed positively or negatively to the overall results. As an example, let's assume that we had a primary meta-analysis addressing the effects of a given intervention on a given outcome encompassing any rodent-based study. At stratified analyses I might want to analyze in separate the effects in studies that have used rats, mice, guinea pigs, gerbils, etc. Only by means of such comparative graphs one can put the data side by side.

Additionally, these comparative graphs have been used in recent preclinical meta-analysis, often associated with forest plots, as can be seen on Sadigh-Eteghad et al. (PLoS One. 2017;12(8):e0184122), Flynn et al. (Front Neurol. 2017;8:357), Chen et al. (PLoS One. 2016;11(7):e0158240), Laban et al. (J Cereb Blood Flow Metab. 2015;35(7):1085-9.) and Hirst et al (Evid Based Preclin Med. 2014;1(1):e00006).
Thus, based on our conception of these comparative graphs and on its recent successful use, most of us prefer to stick with our original decision.

Discussion:
- "the effects observed in animals is…". Should be "are".
Changed, as requested.

- "Showing that lack of sleep leads to a decrease…". This sentence is a fragment.
Fixed, as requested.

- "Both spontaneous and experimental sleep deprivation always coincides…". Should be "coincide".
Fixed, as requested.

- The "Profile of scientific output in the field" seems like an interesting benefit of the study, but the authors never mentioned in the analysis section of the protocol that they would analyze the results by geographical region or year of publication. If they indeed plan to do this, it should be included in the data analysis section.
Originally, our intention was to analyze these factors only narratively; to describe the temporal evolution of the number of articles published in the field, as well as the countries that have published them. However, we agree we should describe how this will be done. We now include this information on our "Data synthesis and analysis" section (pages 12-13).

We would like to thank this Reviewer for his/her careful reading and for the suggestions made to our manuscript. His/her contributions are very interesting and will help to apply the best possible methodological approaches in our systematic reviews.

Reviewer 2:
Thank you for the opportunity to review this systematic review protocol. I have some comments:
1. The Syrcle group have themselves shown the limited utility of funnel plotting in the context of SMD analysis where the weighting is based on variance. This needs to be better justified.
We appreciate Reviewer's comment on this topic and his/her knowledge on recent advances on meta-analytic methods. We do agree that the funnel plot as we have initially planned might lead to biased visual conclusions. We still believe SMD is a better choice for the data we might extract, though. Thus, rather than changing from SMD to NMD, we prefer to stick with SMD and build our funnel plot comparing effect size against the inverse of the squared sample size ($1/\sqrt{n}$), which might generate a better funnel for visual analysis (page 11).

2. Avoidance of the non-peer reviewed literature is questionable, especially if you expect to see publication bias, and given the emerging use of BioRxiv by the community.
This is a very important point to be discussed and we thank the Reviewer for his/her attention on such detailed methodological aspects of our review. As the Reviewer claims, BioRxiv is an emerging tool, reflected by researchers working on preclinical meta-analyses still not reaching a conclusion on how to deal with it. CAMARADES's researchers e.g. have recently published a paper on meta-analytic

methodology with BioRxiv (Wang et al BioRxiv 2018). Neither SYRCLE or CAMARADES has published any position statement or advice about how to proceed.
Not including BioRxiv will increase our risk of publication bias. However, if we would include more than only peer-reviewed literature, considering only BioRxiv is not enough, since many pre-print repositories and similar initiatives continue to be launched (e.g. RIO, PEER J preprint, other journals' preprint). Articles deposited in BioRxiv should eventually be published in a peer-reviewed source. Consequently, not including BioRxiv on our search strategy will work as if we had a publication date cap; if our meta-analysis is updated about two years from now, it should encompass the articles that are today in BioRxiv.
We acknowledge BioRxiv as a very interesting tool for researchers and support its aims. However, we hope to have made clear that there is no consensus on how to use these tools for SRs yet. Thus, we stick to our previous strategy, limiting ourselves to peer-reviewed literature.

3. The authors may wish to reflect on the statistical power of partitioning heterogeneity in SMD analyses in the context of Wang et al BioRxiv 2018 ... unless they have large numbers of studies, they are likely to have limited power. It is what it is, but would be reasonable to discuss.
The referred BioRxiv paper is an important contribution to meta-analytic methodology. It deals with one of the biggest issues in preclinical meta-analysis, which is deciding between the use of either SMD and NMD.
That said, we agree and acknowledge that NMD can provide better estimates than SMD when both are possible and reliably calculated. However, there are cases in which NMD does not perform well, relating to the precision and stability of data acquired for the control groups in each article. Since our samples will probably encompass different behavioral protocols, with results acquired on different scales and with different effect magnitudes, we think that NMD may not work well. Wang's article does not answer whether it is safe to use NMD in comparable cases. Thus, based on our prior experience, we now stick with SMD. We hope that future guidelines will be more precise in stating when NMD and SMD should be used.

4. There could be more definite statement of the stratifications, and what "where possible" means, and what the critical alpha will be for these comparisons (ie are you going to correct for 7 contrasts.
Our intention is to apply the same criteria that drive our decision to perform the first level meta-analyses for the stratified ones. By whenever possible we mean "whenever there are more than three comparable articles". This has been corrected in the manuscript (page 12). Regarding p-values for the stratified analyses, we will apply a Bonferroni correction for the number of analyses to be performed, depending on the nature and amount of records to be included (page 12).

5. I half - get the difficulty in blinding outcome assessment, but there are still ways in which some of the experiment (data analysis, exclusion of outliers etc) might be blinded. And I dont understand why randomisation isn't a thing. And ref 27 is a review of human studies which includes an item on baseline sleep habits ... and is psuedoreplication not an issue in this space, if sleep deprivation exposures are at the level of the cage? So I would articulate, here, the risk of bias items, rather than having the reader track back
Regarding blinding, we have reconsidered a few issues. Blinding as a performance bias item is assessed by the following question: "Were the caregivers and/or investigators blinded from knowledge which intervention each animal received during the experiment?" This kind of blinding is methodologically impossible because any caregiver handling the animals will immediately recognize those housed in standard conditions and those housed in sleep deprivation apparatuses. Blinding in outcome detection is assessed by the following questions "Was the outcome assessor blinded?" In this case, despite we still think any sleep deprivation can be easily recognized by animals body characteristics (posture, sleep pressure, shut eyes, fur, etc), we do acknowledge it is methodologically possible to blind outcome assessors. Thus, we decided to stick with the decision to

remove the blinding on performance out of our list, but blinding in outcome assessment will be incorporated back into our risk of bias analysis.
Regarding randomization, there are two random-related items to be evaluated on a risk of bias assessment: sequence generation (selection bias) and random housing (performance bias). According to the Syrcle risk of bias assessment tool, sequence generation is assessed by the following question: “Was the allocation sequence adequately generated and applied?” It basically inquires whether animals were randomly allocated to the groups. This item will be performed in our analysis, since sleep deprivation protocols do not impact the ability to randomize groups. Random housing is assessed by the question “Was the allocation sequence adequately generated and applied?”. Despite in many cases this is not technically possible, since paradoxical sleep deprivation is often conducted in large water tanks instead of regular home-cages, we do acknowledge it might be possible for some types of sleep deprivation protocols. Thus, this item was incorporated back into our protocol (page 11).
Regarding the use of reference 27, this was a mistake and our intention was to use reference 23 (Pires et al. Neurosci Biobehav Rev. 2016;68:575-89).This was corrected (page 11).
We do acknowledge the risk of pseudoreplication as a issue here, despite it may be minimal. Some (not all) sleep deprivation procedures implies in housing animal on cages containing specific procedural modifications (such as platforms over a water layer, in platform-based paradoxical sleep deprivation methods), but that does not mean that the animals will be subjected to an group-exposure. Also, the sleep deprivation apparatus is exactly the same for every cage, i.e. there are not “levels” of sleep deprivation depending on which cage the animals are housed. Finally, once in the sleep deprivation apparatus, each animal behaves and interacts as a single unit, not as a composite, responding to their own sleep pressure, social hierarchy, etc. These characteristics makes possible to consider each animal as an independent analytical unit, avoid any kind of “cage-bias” and protect the experimental design of potential pseudoreplication.

We are grateful for this Reviewer’s careful and detailed reading. His/her knowledge on meta-analytic methodology is impressive and the suggestions made leaded us to improve our protocol. It will certainly result in more precise analysis when these systematic reviews are performed.

**VERSION 2 - Review**

| | |
|---|---|
| **REVIEWER 1** | ***Olavo Amaral***<br>***Instituto de Bioquímica Médica***<br>***Please state any competing interests or state 'None declared': None declared*** |
| **REVIEW RETURNED** | 23-02-18 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have addressed or justified almost all the issues that were pointed out in the first review of the protocol. A few minor points remain, however:<br>- The authors now include stratified and sensitivity analysis based on research group, which is an interesting addition to the protocol. Nevertheless, they do not define how what will define a research group in their analysis. What exactly makes a set of articles belong to the same group: a single author? The corresponding author? The institution of origin? The possibilities are many and there is no consensus in the published literature on how this should be defined, so it should be stated in the protocol. In case the authors are interested in our graph-based approach (mentioned previously), we will probably have a preprint describing it in a couple of months. We are happy to share code before that, though, and the authors can feel free to contact me so I can put them in touch with the person working on it.<br>- The choice of using stratified analyses instead of meta-regression for quantitative variables does not seem well justified in the authors' responses. If there are continuous variables that have not been collected yet (for which the distribution is still unknown), planning a stratified analysis beforehand may lead one to have a large number of subgroups that could be more easily be described by a single quantitative variable. Besides, pooling multiple values into a few categories will ignore the variation of the impact on heterogeneity within these pools and will include some arbitrary decisions (that should be well described, in case this is indeed chosen). Finally, as shown by Wang et al. (bioRxiv, 2018), stratified meta-analysis to detect moderators can lead to an unacceptable number of false-positives when normalized mean differences are used instead of standardized ones (as the authors plan to use raw mean differences for some analyses, I am not sure if such a problem might arise as well in this case). Using standardized mean differences solves this issue, but leads to lower statistical power than what can be achieved by meta-regression with normalized mean differences (and possibly with raw mean differences as well). Overall, I get the feeling that, at least for quantitative variables, there is no reason to choose stratified analyses over meta-regression to analyse the impact of moderators.<br>- The argument for choosing bar graphs over forest plots also does not seem to hold very well in my view. A forest plot reordered by the variable of interest will have much more visual information on the contribution of each study to the overall effect estimate and heterogeneity. An additional column can inform on the tested variable, which could be the different species, as mentioned. In this case, all experiments using gerbils will come together on top, followed by guinea pigs, mice, rats, and so. Results on the means of each of these subgroups can be added between them, showing the same information as bar graphs, but additionally including information on individual studies as well. Nevertheless, this is also not an impediment for publication. |

| | |
|---|---|
| | - Additionally, I would add small corrections to the following sentences:<br>On page 3, "…on motivated social behaviors such as, aggressive, sexual and maternal behaviors." should be "…on motivated social behaviors, such as aggressive sexual and maternal behaviors." (the authors seem to have deleted the wrong comma when modifying the sentence in response to my comment).<br>On page 3, "…has been focus of research…" should be "…has been the focus of research…"<br>On page 6, "truncated using the wildcard symbol "*'…" has a double quotation mark on one side of the asterisks and single quotation mark on the other. Please correct.<br>On page 10, "…following by assuming the bars..." should be "…followed by assuming the bars…"<br>On page 11, "…and reported the same animal model characteristics…" should be "…and report the same animal model characteristics…"<br>On page 12, "Stratified and sensitive analysis..." should be "Stratified and sensitivity analysis…<br>On page 12, "…number of analysis to be performed." Should be "number of analyses to be performed." |

| | |
|---|---|
| **REVIEWER 1** | ***Malcolm Macleod***<br>***University of Edinburgh***<br>***Please state any competing interests or state 'None declared': I collaborrate with the SYRCLE group*** |
| **REVIEW RETURNED** | 27-02-18 |

| | |
|---|---|
| **GENERAL COMMENTS** | The authors have articulated the reasons for the analysis choices which they have made. |

**VERSION 2 – AUTHOR RESPONSE**

The authors have addressed or justified almost all the issues that were pointed out in the first review of the protocol. A few minor points remain, however:

• The authors now include stratified and sensitivity analysis based on research group, which is an interesting addition to the protocol. Nevertheless, they do not define how what will define a research group in their analysis. What exactly makes a set of articles belong to the same group: a single author? The corresponding author? The institution of origin? The possibilities are many and there is no consensus in the published literature on how this should be defined, so it should be stated in the protocol. In case the authors are interested in our graph-based approach (mentioned previously), we will probably have a preprint describing it in a couple of months. We are happy to share code before that, though, and the authors can feel free to contact me so I can put them in touch with the person working on it.

o The offer for sharing the methodology to perform the aforementioned graph-based approach is much appreciated. However, as these research group analyses are a secondary aim in this manuscript, we prefer to stick with a simpler and more straightforward analysis at this time. By our familiarity with the field, we know that there are no more than four research groups working consistently on the relationship between sleep deprivation and maternal behavior, for instance. Thus, in this case, a more complicated method would not add much. We do agree that our methods should be better described, though. A short paragraph has been added to make clear how research groups will be defined and analyzed (page 12). Regarding the graph-based analysis, we would be glad to receive a pre-print for other ongoing and future work, and will encourage it being applied on a larger scale.

• The choice of using stratified analyses instead of meta-regression for quantitative variables does not seem well justified in the authors' responses. If there are continuous variables that have not been collected yet (for which the distribution is still unknown), planning a stratified analysis beforehand may lead one to have a large number of subgroups that could be more easily be described by a single quantitative variable. Besides, pooling multiple values into a few categories will ignore the variation of the impact on heterogeneity within these pools and will include some arbitrary decisions (that should be well described, in case this is indeed chosen). Finally, as shown by Wang et al. (bioRxiv, 2018), stratified meta-analysis to detect moderators can lead to an unacceptable number of false-positives when normalized mean differences are used instead of standardized ones (as the authors plan to use raw mean differences for some analyses, I am not sure if such a problem might arise as well in this case). Using standardized mean differences solves this issue, but leads to lower statistical power than what can be achieved by meta-regression with normalized mean differences (and possibly with raw mean differences as well). Overall, I get the feeling that, at least for quantitative variables, there is no reason to choose stratified analyses over meta-regression to analyze the impact of moderators.
o While we agree with the reviewer regarding the statistical power and advantages of meta-regression over stratified analysis, we based our decision on more than only statistics. We know meta-analyses of animal data encompass high heterogeneity and uncertain impact of biases on the results, consequences of some unfortunate intrinsic characteristics of currently published animal research, such as the low quality on the reports. Guidelines on the performance of preclinical meta-analysis clearly suggest not to exclude low-quality articles, but rather to include as much sources as possible in order to be able to reach some results (if we would exclude them all as we do in clinical meta-analysis, we would probably end up with nothing in hands to analyze). Additionally, in contrast with clinical meta-analysis, which are much more oriented to estimate effects, preclinical meta-analysis are conceptually more applicable as an exploratory and hypothesis-generating tool (Hooijmans et al., 2014). Considering this, most of us think stratified analysis, combined with sensitivity analysis, might be better to explore data and provide insights regarding future studies, methodological caveats and translational potential. That said, we prefer to stick with our decision to perform stratified analysis rather than meta-regressions.

• The argument for choosing bar graphs over forest plots also does not seem to hold very well in my view. A forest plot reordered by the variable of interest will have much more visual information on the contribution of each study to the overall effect estimate and heterogeneity. An additional column can inform on the tested variable, which could be the different species, as mentioned. In this case, all experiments using gerbils will come together on top, followed by guinea pigs, mice, rats, and so. Results on the means of each of these subgroups can be added between them, showing the same information as bar graphs, but additionally including information on individual studies as well. Nevertheless, this is also not an impediment for publication.
o We do acknowledge forest plots are far more common and accepted for visualization of meta-analytic data, but as previously argued we think comparative bar graphs have their advantages as well. In any case, this is a matter of preference, since the calculations and interpretations do not depend on how the data are displayed. These comparative bar charts have been consistently used by the Camarades group and have also been used in a previous article of our group (Pires et al., 2016). The main benefit in favor of comparative graphs is that they summarize information while forest plots would be substantially longer. For example figure 3 in Pires et al., 2016 presented as a series of forest plots would have more than 200 entries. We will however create forest plots for all comparisons and include them in our publication as supplementary files (page 13).
• Additionally, I would add small corrections to the following sentences: On page 3, "…on motivated social behaviors such as, aggressive, sexual and maternal behaviors." should be "…on motivated social behaviors, such as aggressive sexual and maternal behaviors." (the authors seem to have deleted the wrong comma when modifying the sentence in response to my comment).
o Corrected, as requested.

• On page 3, “…has been focus of research…” should be “…has been the focus of research…"
o Corrected, as requested.
• On page 6, “truncated using the wildcard symbol “*’…” has a double quotation mark on one side of the asterisks and single quotation mark on the other. Please correct.
o Corrected, as requested.
• On page 10, “…following by assuming the bars...” should be “…followed by assuming the bars…”
o Corrected, as requested.
• On page 11, “…and reported the same animal model characteristics…” should be “…and report the same animal model characteristics…”
o Corrected, as requested.
• On page 12, “Stratified and sensitive analysis...” should be “Stratified and sensitivity analysis…
o Corrected, as requested.
• On page 12, “…number of analysis to be performed.” Should be “number of analyses to be performed.”
o Corrected, as requested.

We would like to thank this Reviewer for his careful reading and suggestions. His knowledge on preclinical meta-analysis is impressive. Despite us not agreeing with all aspects, this discussion has been important to improve this protocol even more. We are grateful for all his attention.

#2 Submitted by : Malcolm Macleod
• The authors have articulated the reasons for the analysis choices which they have made.

We are grateful for this Reviewer’s reading and approval. Having the consent of such a reputable researcher is an honor. His advice during the previous revision round was of major importance and has helped us to improve this protocol.